Journal of Attention Disorders Vol. 12(1):15-43 (2008)
ISSN: 1087-0547
doi: 10.1177/1087054708319525

Evidence, Interpretation, and Qualification From Multiple Reports of Long-Term Outcomes in the Multimodal Treatment Study of Children With ADHD (MTA) Part II: Supporting Details

James Swanson, L. Eugene Arnold, Helena Kraemer, Lily Hechtman, Brooke Molina, Stephen Hinshaw, Benedetto Vitiello, Peter Jensen, Ken Steinhoff, Marc Lerner, Laurence Greenhill, Howard Abikoff, Karen Wells, Jeffery Epstein, Glen Elliott, Jeffrey Newcorn, Betsy Hoza, and Timothy Wigal. ***MTA Cooperative Group***

Objective:

To review and provide details about the primary and secondary findings from the Multimodal Treatment study of ADHD (MTA) published during the past decade as three sets of articles.

Method:

In the second of a two part article, we provide additional background and detail required by the complexity of the MTA to address confusion and controversy about the findings outlined in part I (the Executive Summary).

Results:

We present details about the gold standard used to produce scientific evidence, the randomized clinical trial (RCT), which we applied to evaluate the long-term effects of two well-established unimodal treatments, Medication Management (MedMGT) and behavior therapy (Beh), the multimodal combination (Comb), and treatment "as usual" in the community (CC). For each of the first three assessment points defined by RCT methods and included in intent-to-treat analyses, we discuss our definition of evidence from the MTA, interpretation of the serial presentations of findings at each assessment point with a different definition of long-term varying from weeks to years, and qualification of the interim conclusions about long-term effects of treatments for ADHD based on many exploratory analyses described in additional published articles.

Conclusions:

Using a question and answer format, we discuss the possible clinical relevance of the MTA and present some practical suggestions based on current knowledge and uncertainties facing families, clinicians, and investigators regarding the long-term use of stimulant medication and behavioral therapy in the treatment of children with ADHD. *(J. of Att. Dis. 2008; 12(1) 15-43)*

A. Introduction, Purpose, and Background for Part II (Supporting Details)

1. Introduction

Integrating the main findings of the series of three main reports (consisting of 8 articles) is not a simple task. Presenting the findings from multiple articles at each assessment points is complex, and linking these serially over the three assessment points magnifies the complexity. The interpretation and qualification of the main findings that incorporate information from the

many additional articles (60 or more) further increases the complexity of this task. We thought a detail-heavy presentation might be distracting to some readers, but we anticipated that other readers would appreciate details about the evidence, interpretation, and qualification of the findings. Because of this, we decided to prepare a two-part article: in Part I (Executive Summary) we outlined issues without excessive detail, and here in Part II (Supporting Details) we will provide a plethora of details.

2. Purpose

Our purpose is to provide extensive (and perhaps excruciating) details necessary to explain why we believe that (a) the primary articles provide the main evidence from the MTA and must be narrowly focused, (b) the secondary articles provide broader interpretations that should be used to supplement but not replace this main evidence, and (c) the many additional articles (see Appendix) provide qualifications that should be considered exploratory, not definite, and used to generate new hypotheses to be tested in future research. In the final section on Clinical Implications, our purpose is to provide more details about future directions suggested by the interim findings of the MTA in a "question and answer" format.

3. Background

a. Genesis of the MTA

The MTA was conceived at the beginning of an era (1990 to 2000) when the budget for NIH was increasing and public funding for large groundbreaking projects was feasible. This preceded the proliferation of industry funded studies that emerged in the late 1990s after the Food and Drug Modernization Act created financial incentives (the "pediatric exclusivity" rule) for pharmaceutical companies to evaluate medications in children (see DeVeaugh-Geiss et al., 2006). The objectives of the MTA (see Richters et al., 1995) were described in a detailed Request for Applications (RFA) developed by the National Institute of Mental Health (NIMH) in the early 1990s. This RFA identified a need for long-term studies of established treatments for ADHD. Information from many short-term studies provided support for the clinical use of two treatment modalities, which we labeled medication management (MedMgt) and behavior therapy (Beh), but few long-term studies had been conducted of either of these or their combination. The RFA called for an ambitious long-term study to fill this gap.

The review process of the proposals submitted in response to the RFA led to the selection of six sites for the study. Each of the proposals had a different specific design for the requested prospective study of long-term effects of treatments of ADHD, so it was necessary to develop consensus among the investigators from the multiple sites and the NIMH staff leading this initiative, which we called the MTA Cooperative Group. Our initial task was to develop a common protocol for the MTA. The evolution of the design and methods of the MTA were described in the initial publications by the MTA Cooperative Group in the late 1990s about the medication algorithm (Greenhill et al., 1996), the design (Arnold et al., 1997a, 1997b), and the assessment battery (Hinshaw et al., 1997), and later about the psychosocial treatments (Wells et al., 2000a, 2000b).

b. Design and Analysis Framework for the MTA

Early in the design process, we embraced the logic and methodology of the gold standard (see Kraemer & Robinson, 2005) for evidence-based medicine—the randomized clinical trial (RCT). As described by Gibbons et al. (1993), this requires that participants be randomly

assigned to the treatment conditions to ensure the formation of groups would not be biased by preferences of those providing or receiving the treatments or by other factors that may operate intentionally or unintentionally. If the sample size is large enough, randomization should produce groups that do not differ in any systematic way at the baseline assessment. As part of the design phase of the MTA, a statistical power analysis was used to estimate the required sample size (n = 576 to provide 144 participants per treatment group). During the recruitment and enrollment phase, the sample size requirement was exceeded by 3 cases, and n = 579 cases were diagnosed and randomized to the treatment conditions, with at least n = 24 per site assigned to each of the 4 groups to allow for tests of site by treatment interactions (see Kraemer and Robinson, 2005). The RCT approach dictates the use of intent-to-treat (ITT) analyses, in which all participants who are randomized are analyzed to evaluate the effects of the assigned treatments, which may (and often do) differ from the actual treatment received by participants in the study. The actual treatment depends on whether the assigned treatments are accepted and implemented as intended. ITT analyses are intended to evaluate the recommendation for treatment and thus do not take actual treatment into account, although secondary analyses may do so to supplement the primary analyses.

Also, early in the design process, we adopted a statistical method called based on random regression models (see Gibbons et al., 1993). This technique evaluates the change over time by estimating for each participant the intercept and slope of a regression equation relating outcome to time. By averaging these values across participants in each group, estimates are provided for the slope and intercept for the treatment. Based on random assignment, the treatment groups are not expected to differ on any measure obtained from the baseline assessment. If this is considered time = 0, then it defines the intercept of the regression equation for that measure. However, if the treatments result in differential change over time, then differences in the slope of the regression equation are expected. In the RCT approach, this type of regression analysis is applied to estimate relative rather than absolute effects of treatment (see Gibbons et al., 1993; Kraemer, Wilson, Fairburn, & Agras, 2002). Relative treatment effects are tested by contrasting slopes of regression lines describing outcome over time for the groups, which in our analyses could be accomplished by the evaluation of the treatment × time interaction.

c. Number, Intensity, and Duration of Treatments

The RFA directed us to select and contrast established treatments. In the initial meetings of the MTA Cooperative Group, a consensus was reached to evaluate two modalities of treatment, medication management (MedMgt) and behavior therapy (Beh), offered either alone or in combination (Comb). These treatment conditions were to be contrasted not only with each other but also with a fourth treatment condition that we called Community Comparison (CC). Because the short-term effectiveness of the two treatment modalities had been documented, we did not consider it ethical to insist on a true "no-treatment" control group in a long-term intervention study. Also, we thought it would be impractical to deliver a sham or placebo treatment over a long duration. Therefore, we decided to use a "treatment-as-usual" contrast group that we considered to be ecologically valid and consistent with the aims of our RCT. Some have agreed with us (see Barkley, 2000) but others have not (see Breggin, 2001). We used the word *comparison* instead of *control* because we expected participants in this group to receive a variety of treatments in the community, including treatment with stimulant medications.

Decisions were required about the intensity of treatment and how long treatment should be provided. One of the first controversies our MTA Cooperative Group faced was about the intensity of treatment. We decided to provide state-of-the-art treatments, which made the MTA

an efficacy study of very intensive treatments generally unavailable at the time rather than an effectiveness study of available treatments. A medication algorithm was developed for the MTA (see Greenhill et al., 1996) based on treatment with the most frequently prescribed stimulant medication, which in 1994 was immediate-release methylphenidate. To achieve coverage across the day, 3-times-a-day dosing was used, and to achieve coverage over the long term, medication was administered without weekend or summer holidays (i.e., 7 days a week, 365 days a year). A 4-week double-blind titration was used to evaluate responses to a range of doses (from placebo to 20 mg) and to select the best per-administration dose for each participant (see Greenhill et al., 2001). After titration, monthly clinic visits were scheduled for close monitoring by the MTA pharmacotherapists, with sufficient time allowed (i.e., 30 minutes) to discuss issues about attitude toward and adherence to the medication regime. The protocol required them to obtain information from the child's teacher as well as from parents and from the child (in absence of the parents). Based on these sources, regular adjustments of dose were made (as needed) to maintain maximum benefit (see Vitiello et al., 2001).

The MTA behavioral therapy was similarly intensive (see Wells et al., 2000a, 2000b). It consisted of three behavioral components: (a) a 27-session group parent training intervention supplemented with 8 individual sessions during the 14-month treatment phase; (b) a school intervention spanning 2 school years and including teacher consultation and a trained paraprofessional aide in the classroom of every participant for 12 weeks; and (c) an 8-week Summer Treatment Program. The components were coordinated over the 14-month treatment phase by the same behavior therapist for each participant, who conducted both the parent training and teacher consultation components that were faded over time.

Based on our review of the literature on pharmacological treatments (see Greenhill et al., 1996), we expected that in most cases the medication management delivered according to clinical practices in the community (i.e., number of clinic visits, doses of medication evaluated, sources of information about responses to treatment, etc.) would vary, but on the average we expected this treatment-as-usual would be less intensive than in our state-of-the-art medication management according to the MTA medication algorithm. Based on our review of the literature on psychosocial treatments (see Wells et al., 2009), we also expected that behavior therapy would not be generally available in the community, but if it were available, we expected the intensity of this treatment-as-usual would vary, too. As with medication, we expected on the average this would be less than in our state-of-the-art MTA behavioral algorithm. Based on the expectation of varying percentage of participants receiving treatment and varying intensities of those treatments in the communities where the study was to be implemented, we realized that the CC group might have greater variance on the outcome measures sensitive to treatment effects. This is a common assumption about treatment-as-usual compared to treatment-by-protocol, which has implications for the statistical comparisons that include this group (see NIH Conference on Considering Usual Medical Care in Clinical Trial Design, 2005). If the variance of the CC group were greater than the other groups, this could adversely affect statistical comparisons involving this group: For a given difference between average outcome for the CC and another group, the statistical significance would be reduced, unless a large sample was used to compensate for the hypothesized larger variance. (In fact, the variances for the most sensitive outcome measures, ADHD symptom ratings, were not larger in the CC group compared to the other groups [see Swanson, 2005]).

During the design phase we had to set the length of our treatment period and the timing of our assessments. We expected the effects of one modality (treatment with medication) to

emerge more rapidly than effects of the other modality (treatment with behavior therapy), and therefore we expected that the relative effects—the differences in outcome for the two treatment modalities— would vary over time. We decided that the most critical comparison would be at the end of the treatment phase, which we wanted to be as long as possible. Although we initially hoped to extend active treatment for 2 years, we were limited by the available budget for the MTA. Thus, we set the duration of treatment to be slightly over a year (14 months) to cover 2 school years and the summer between them (see Wells et al., 2000). To address interim changes in symptoms and impairment, we scheduled 2 assessments at 3 months and 9 months after baseline, in addition to the end-of-treatment assessment at 14 months after baseline.

d. Randomization and Selection Bias

All children in the MTA met the same entry criteria (which included confirmation of a diagnosis of ADHD-Combined Type), but there was variation across these cases in severity of the symptoms of ADHD, presence of comorbid disorders, family characteristics, and many other factors. The process of randomization provides an unbiased assignment of participants to the treatments and thus protects against confounding these or others characteristics with the treatment conditions (see Gibbons et al., 1993). The general concern is that treatment determined by choice (of the clinician or the participant) might result in the tendency for the more severely affected cases to receive one type of treatment (e.g., stimulant medication) and the less severely affected cases another type of treatment (e.g., behavior therapy). When this occurs, the group choosing the preferred treatment (even if it is more effective) may have worse outcomes than the group that received the less preferred treatment, because groups determined by choice rather than randomization may differ in severity from the start—leading to different outcomes over time. In short, random assignment protects against an effect known as the intervention-selection bias (e.g., the most severely affected receiving the most intensive treatment, leading to an apparent association of such treatment with the worst outcome).

The randomly assigned treatment may differ from the actual treatment received by participants, because acceptance of assigned treatment could be rejected and adherence to the recommended treatment regimes could be low. However, in RCT methodology, the ITT principle dictates that the analyses compare participants on the basis of their assigned groups regardless of adherence. Based on this, we adopted a strict rule for evaluating the participants who entered the study ("once randomized, always analyzed"). There was no switching of groups even if a participant switched treatments (e.g., if a participant assigned to the Comb group accepted the Beh component but refused the MedMgt component, in the ITT analysis the participant was retained in the Comb group and was not switched to the Beh group even though that was the actual treatment received). Thus, ITT analyses evaluate the recommended treatment rather than the actual treatment received.

The ITT principle does not preclude observing actual treatment and using this information in secondary analyses. Of course, the secondary analyses of self-selected subgroups (e.g., based on acceptance and participation) are not protected by randomization from confounding effects of known or unknown factors that might be correlated with subgroup membership. However, these analyses are important to evaluate another potential selection bias that may operate in a RCT, which may be called the participation-selection bias (see Marcus & Gibbons, 2001). A participation-selection bias may result in an underestimate of the effect of a treatment in an ITT analysis when not all participants assigned the treatment actually receive it. We developed the Services for Children and Adolescents–Parent Interview (SCAPI) to track information about the actual treatment received, based on participant choice and ability to obtain

services in the community (see Hoagwood et al., 2004; Jensen et al., 2000). We used the information from the SCAPI at each assessment point to establish patterns of medication use (i.e., consistently treated or inconsistently treated, never treated or newly treated, stopping treatment or starting treatment, etc.). These patterns defined naturalistic subgroups of participants, which were used in secondary analyses to supplement (but not replace) the primary analyses.

e. Multiple Outcome Measures

An important decision during the design process concerned the choice of outcome measures to be evaluated and tested in specific a priori hypotheses. Some members of the MTA cooperative group were proponents of a single global outcome measure, whereas others preferred multiple outcome measures from the large assessment battery (see Hinshaw et al., 1997) to evaluate differential impact of the treatment conditions on various outcome domains. Indeed, because ADHD symptoms per se may not be the best predictors of ultimate outcome, we decided to use multiple outcome measures, even though this procedure required adjustment for multiple comparisons in analyses of outcome to protect against false positive effects. At each assessment point, our plan was to repeat the full assessment battery with multiple measures from multiple methods of assessment and multiple sources of information. In the complete set of measures we expected redundancy, so we planned to use data reduction methods to determine a smaller set from different sources (e.g., parent, teacher, and child ratings) and various domains (e.g., ADHD and non-ADHD symptom domains, areas of functional impairment).

For statistical evaluation of treatment effects, the convention is to set a significance level (e.g., $p < .05$) to limit false positive findings. To maintain this level of significance in the face of multiple comparisons, an adjustment of the significance level is necessary to keep the overall chance of a false positive at $p < .05$. A common correction called the Bonferroni adjustment is made simply by dividing the significance level by the number of comparisons. We were aware that the use of multiple outcome measures would result in a more stringent significance level for each of the statistical tests than if we used just one outcome measure. However, the complexity of ADHD and the importance of considering multiple domains of impairment as well as symptoms dictated our decision to use multiple outcome measures. This clearly contributed to confusion and controversy, which likely would have been much less had we specified one outcome measure as primary and explicitly considered all other uncorrelated outcome measures as secondary to merely elucidate the results on the primary.

For the initial evaluation of long-term effects, our procedures identified six distinct domains—ADHD symptoms, oppositional/aggressive symptoms, internalizing symptoms, social skills, parent-child interactions, and academic achievement—with a total of 19 marker outcome measures (MTA Cooperative Group, 1999a). This provided the basis for adjustment of the significance level of the MTA (p .05) by domain (.05/6) or by measure (.05/19) or in a sequential analysis by some combination of the domains and measures (see below).

f. Multiple Comparisons

As part of the design phase of the MTA, we formulated primary questions based on comparisons of any two treatment conditions in terms of change in outcome measures over time. Thus, we evaluated the effects of the four randomly assigned treatment conditions (average slope of the regression curves) compared to each other (i.e., relative effects) rather than the absolute effects compared to a no-treatment or placebo control condition that was not used in the MTA.

With four treatment groups there are six paired-comparisons. Power analyses were performed taking the multiple sites into consideration (see Kraemer & Jackson, 2005). Under the assumptions of an effect size = 0.4 and power = 0.81, 24 participants per treatment per site (a total of 144 per treatment across sites) were required for the six possible comparisons of any two treatment conditions on the primary outcome measure of overall rating of ADHD core symptoms (see MTA Cooperative Group, 1999a). We organized these into three sets, with some directional hypotheses evaluated by one-tailed tests (denoted by "<" or ">") and nondirectional hypotheses evaluated by two-tailed tests (denoted by "-"). The cutoff for statistical significance is more stringent for a two-tailed than a one-tailed test. The first set consisted of a nondirectional comparison of the unimodal treatments (MedMgt - Beh). The second set consisted of two tests based on directional comparisons of multimodal treatment to each of the unimodal treatments, with the expectation of multimodal superiority (Comb > MedMgt and Comb > Beh). The third set consisted of three directional comparisons of community treatment (treatment-as-usual) to the other treatments provided by MTA staff (treatment-by-protocol). Our treatment-as-usual condition (CC) was expected to be inferior to each treatment-by-protocol condition (CC < Comb, CC < MedMgt, and CC < Beh).

Even though there are six possible pairwise comparisons for the four groups, only three comparisons can be used that will be statistically independent of each other. It is possible to construct several different sets of three contrasts that will be independent using the method called orthogonal contrasts. One set was proposed by Swanson et al. (2001): (a) the MTA Medication Algorithm contrast by comparing the average of two groups, both including the MTA method of intensive medication management as a component of the assigned treatment, to the average of the other two groups [(Comb + MedMgt) > (Beh + CC)]; (b) the Multimodal Superiority contrast that was the same as one of the pairwise comparisons (Comb > MedMgt); and (c) the Psychosocial Substitution contrast by comparing the generally unavailable intensive behavioral treatment of the MTA to generally available community standard treatment that in most cases included stimulant medication (Beh > CC).

g. Adjustment for Significance Level

In the primary analyses at the end of treatment (MTA Cooperative Group, 1999a), we initially planned to evaluate the overall rating of ADHD core symptoms as the primary outcome measure, and for the method of multiple comparisons this would require dividing the significance level by the number of comparisons (.05/6 = 0.008). But, to make 6 comparisons for 19 outcome measures, the significance level should be divided by their product [e.g., .05/ (19 × 6)]. This would require a drastic adjustment of the significance level from $p \leq .05$ to $p < .05/114 = .0004$. However, we decided to perform sequential tests to justify a less severe adjustment. First, within each domain we adjusted our significance level based on the number of outcome measures within that domain (e.g., for the five outcome measures in the ADHD symptom domain, our adjustment was .05/5 = .01). Second, the six multiple comparisons for a given measure within a domain were made only if the omnibus test of the treatment × time interaction was significant, and for these the omnibus-adjusted significance level was further adjusted by dividing by 6 (i.e., for any significant measure within the ADHD domain, .01/6 = .002). These adjustments yielded significance levels of $p \leq .002$ to $p \leq .004$, depending on the number of measures within each domain (see MTA Cooperative Group, 1999a).

h. Multiple Definitions of Long-Term

The fundamental purpose of the MTA was to evaluate long-term effects of treatment. In the early 1990s a large literature existed that documented clear short-term beneficial effects (i.e., reduction of severity of ADHD symptoms) of stimulant medication and behavior therapy. The MTA started with documentation of short-term effects as it unfolded in a prospective fashion to evaluate long-term effects. The short-term effects of MedMgt were evaluated in a double-blind 30-day titration trial, which verified beneficial effects manifested by reduction of ADHD symptoms and related behaviors on the first day, during the first week, and over the first month of treatment (Greenhill et al., 2001). Intermediate effects were estimated from the assessments at 3 months and 9 months after randomization (see Arnold et al., 2004). Also, intermediate effects were provided by a comparison of two treatments (Comb and Beh) during the implementation of one component of behavior therapy, the summer treatment program (Pelham et al., 2000). However, the fundamental purpose of the MTA was to assess the long-term effects of treatment, and for the initial end-of-treatment analyses in our RCT design, long-term was defined by ITT analyses of outcomes at the 14-month assessment.

i. A Priori and Post Hoc Tests

As in any area of investigation in science or medicine, the rigorous evaluation of a limited number of primary hypotheses in the MTA will leave many additional questions unaddressed. For example, at the initiation of the MTA, the hypothesis of stimulant-related growth suppression had been discounted (see Roche et al., 1979). Based on this literature, we did not include the hypothesis of growth suppression in our specific aims, and we had not listed growth suppression as a possible side-effect of treatment with medication in our consent forms. Before our initial findings were published, an additional influential study (Spencer et al., 1996) and another comprehensive review (NIH Consensus Conference on ADHD, 1998) discounted the hypothesis of stimulant-related growth suppression. Based on this, we did not include an evaluation of growth in our initial publication of findings at the end of the treatment phase (MTA Cooperative Group, 1999a). However, because we included the measurement of physical size at each assessment, we conducted the post hoc evaluation of the growth suppression. This analysis was not published until we reported the findings from our first follow-up evaluation (see MTA Cooperative Group, 2004b).

For post hoc tests, both false positive results (related to the many statistical comparisons) and false negative results (related to size of detectable effects given reduced sample size of various subgroups) are expected, so these supplemental reports should not be used to accept or dismiss any specific hypothesis with firmness. The findings of post hoc tests should be interpreted with caution and with full recognition that they generate the next set of hypotheses to consider for rigorous evaluation in future randomized clinical trials (see Kraemer et al., 2002).

B. Publications at the End of Treatment (MTA Cooperative Group, 1999a, 1999b)

1. Evidence: ITT Analyses

Our use of multiple outcome measures and multiple comparisons introduced a level of complexity that has surely contributed to controversies related to the findings from the MTA. We do not have a single test, or even a few tests, to characterize the findings of the MTA. Instead, in our primary analyses at the end of treatment (MTA Cooperative Group, 1999a), the omnibus test of a significant interaction of treatment condition with time was significant for 10 of the 19 outcome measures (4 were from the ADHD symptom domain and 6 were from the other domains). Therefore, we performed $10 \times 6 = 60$ paired-comparisons of the treatment conditions,

and 23 of these were significant at the adjusted significance levels (see MTA Cooperative Group, 1999a, Table 5).

For the contrast of unimodal treatments (MedMgt - Beh), we reported that MedMgt was better than Beh on 3 of the 10 outcome measures that had revealed significance in the omnibus tests, and all of these were ratings of ADHD symptom severity. For the multimodal treatment comparisons (Comb $\succeq$ MedMgt and Comb $\succeq$ Beh), we reported that Comb was not better than MedMgt on any of the 10 outcome measures but was better than Beh on 6 of them (and 3 of these were ratings of ADHD symptom severity). For the comparisons to community treatment (CC $\preceq$ Comb, CC $\preceq$ MedMgt, and CC $\preceq$ Beh), we reported that, of the 10 outcome measure, CC was worse than Comb on 9, MedMgt on 5, and Beh on 0.

The synthesis of these analyses reveals two key findings (see MTA Cooperative Group, 1999a). First, assignment to pharmacological treatment according to the intensive MTA medication algorithm (MedMgt) produced larger benefits (i.e., reduction in ratings of ADHD symptom severity) than long-term psychosocial treatment according to the intensive MTA behavior therapy algorithm (Beh). Second, relative benefit from assignment to the multimodal combination (Comb) was not significantly greater than to the MTA medication algorithm alone.

2. Interpretation: Moderator and Mediator Analyses

Kraemer et al. (2002) and Hinshaw (2007) provide cogent discussions of moderator and mediator analyses in the context of clinical trials. The secondary analyses (MTA Cooperative Group, 1999b) addressed some of these issues by performing moderator analyses of heterogeneity within the randomized groups (e.g., subgroup analyses based on measures from the baseline assessment battery obtained prior to randomization) and mediator analyses of variables related to treatment implementation (e.g., actual treatment related to acceptance of and adherence to the MTA protocols).

In the secondary article at the end of treatment (MTA Cooperative Group, 1999b), moderator analyses were conducted for five variables (sex, prior medication status, comorbid oppositional defiant or conduct disorder, comorbid anxiety disorder, welfare status) and for one mediator variable (acceptance and attendance related to treatment sessions deemed to be as-intended or below-intended). For each of these, random regression analyses were performed for 14 of the 19 outcome measures. This required 84 analyses, but these were considered exploratory analyses for generating hypotheses for future research so the Bonferroni adjustment to the significance level was not used.

The moderator analysis suggested that the pattern of treatment response (i.e., statistical significance on more than one outcome measure) differed for subgroups defined by comorbid anxiety disorders (present or not) and family income (public assistance or not). In the subgroup of the ADHD participants with anxiety disorders at baseline (34% of the sample), Beh was not significantly worse than either MedMgt or Comb on any outcome measure and was significantly better than CC for three outcome measures (parent rating ratings of ADHD symptoms and internalizing social skills and child interview of anxiety symptoms). In the subgroup on public assistance, the relative effect of Comb was superior to all other treatments for two outcome measures (teacher ratings of total social skills and parent ratings of personal closeness with the child).

For mediator analyses, we identified subgroups with as-intended treatment according to the MTA medication algorithm (78%) and the behavior therapy algorithm (63%). Mediator analysis revealed that for the medication component in the MedMgt and Comb treatment conditions, the subgroups with as-intended treatment manifested a greater response (reduction in

ADHD symptoms over time) than the subgroup with below-intended participation in the MedMgt group. Interestingly, for the behavior therapy component of the Beh and Comb treatment conditions, the mediator analyses indicated that the as-intended and below-intended subgroups did not differ. Thus, the lower acceptance and attendance of the Beh than the MedMgt component did not account for the significant difference between these two unimodal treatment conditions in the primary ITT analyses. The acceptance and attendance measures may have been insufficient to serve as mediator variables, and other factors (actual parent and teacher implementation of behavioral programs) that likely would have been superior for this purpose but were not assessed and could not be evaluated as mediators.

In the secondary article (MTA Cooperative Group, 1999b), in addition to the mediator analysis of acceptance and attendance related to the MTA treatment algorithms, these variations of medication use in the community were addressed. The CC subgroup that received treatment with medication in the community (68% of the CC participants) had better outcome than the subgroup that did not. Greenhill et al. (submitted) documented that for the CC group, medication was prescribed at lower doses, the frequency of dosing was less (usually 2 rather than 3 times a day) and was monitored with fewer office visits (roughly 9 minutes every 5-6 months vs. 30 minutes every month), and direct input from the child's teacher was not usually obtained. The medication-treated subgroup of the CC group had better outcome than the nonmedication subgroup, which was similar to the outcome of the Beh group but was worse that the outcome of the MedMgt group. This pattern suggests that the intensive Beh treatment of the MTA and medication treatment-as-usual may have about the same effectiveness (although we should only assert that we could not document a statistical difference), whereas more intensive treatment according to the MTA medication algorithm may have greater efficacy than treatment-as-usual with medication by clinicians in the community.

3. Qualification: Additional Exploratory Analyses

a. Additional Moderator Analyses

Jensen et al. (2001a, 2001b) reported that children with ADHD plus anxiety comorbidities (but not disruptive comorbidities) did relatively well with Beh, but children with ADHD plus both anxiety and disruptive comorbidities had poor outcomes unless they received Comb. For ADHD cases with the double comorbidity of disruptive behavior (e.g., oppositional or conduct disorder) and internalizing disorders (e.g., anxiety or depression), Comb was the only treatment condition leading to meaningful gains. Thus, in the MTA trial, some participants were relatively refractory to the intensive medication management provided by the MTA, revealing that additional components of treatment may be necessary to benefit the cases with the greatest severity of ADHD symptoms and adversity in life.

Owens et al. (2003) reported that even though the MTA medication algorithm was quite effective overall for improving ADHD and ODD symptoms in the sample, it was notably less effective for children who entered the study with severe symptom levels of ADHD, for those whose primary caregivers had even mild levels of depression, and for those with subaverage IQ score.

Arnold et al. (2003) performed an analysis of ethnicity and reported a significant interaction for this moderator. Caucasians showed no advantage of Comb over MedMgt, but pooled minorities did, with an effect size of 0.36.

Hechtman et al. (2005) evaluated another interesting and important outcome measure (e.g., presence of comorbid disorders), and reported that Comb, but not MedMgt, was

statistically superior to CC for reducing the persistence or development of comorbid oppositional-defiant and mood disorders.

Santosh et al. (2005) performed a moderator analysis based on subgroups defined by the rediagnosis of the DSM-IV cases by applying criteria and guidelines for the ICD-10 diagnosis of Hyperkinetic Disorder (which requires greater pervasiveness of symptoms across home and school setting than the DSM-IV diagnosis of ADHD-Combined Type) using a well-developed algorithm ("Hypescheme") to consolidate information from the assessment battery. Only 25% of the MTA cases with a DSM-IV diagnosis of ADHD Combined Type also received an ICD-10 diagnosis of Hyperkinetic Disorder. In the subgroup with Hyperkinetic Disorder, the response to MedMgt was larger and the response to Beh was smaller than in the other subgroup, so the difference between MedMgt and Beh was very large. However, for the remaining 75% of the MTA cases, the difference between MedMgt and Beh was much smaller. The greater relative superiority of medication at the end of the 14-month treatment phase was suggested as evidence in support of the recommendation of medication as the first choice to initiate treatment in cases with Hyperkinetic Disorder. For those without the extreme level of pervasiveness of symptoms across domains and sources required for this diagnosis, Santosh et al. (2005) speculated that nonpharmacological treatment may be preferred as the first choice for treatment, with a provision to add medication later if necessary. This provided some justification for a second option for staging treatment components. In less severe cases, Beh might be considered as a first choice when cultural or national differences (see Taylor, 1999) provide a context of low preference for medication.

We also explored difference among subgroups in responses to behavior therapy. Hinshaw et al. (2000) reported that improvements in parents' negative/ineffective discipline helped to explain major improvements in both school-based social skills and disruptive behavior for youth randomly assigned to Comb. For families who showed meaningful reductions in negative/ineffective discipline styles during the course of the trial, only Comb led to high rates of normalized functioning by the end of 14 months of active intervention. In qualitative analysis of a cutoff based on symptom severity intended to denote successful treatment, the percentage of cases considered to be successfully treated in the Comb group (68%) was significantly higher than in the MedMgt group (56%).

Thus, both the percentage of participants treated with the MTA medication algorithm and the behavioral algorithm and the characteristics of the treatment regimes, varied across the assigned groups and within each treatment group (i.e., in some identifiable subgroups). This complexity contributed to controversy about who was treated in the various subgroups and how they were treated with stimulant medication and with behavior therapy. These selection and implementation factors may affect outcome substantially, but could not be adequately evaluated in the MTA.

b. Multiple Comparisons Across Outcome Domains

The decision to use multiple outcome measures appeared to be justified, because these analyses suggested some key differences across domains. With so many tests, we resorted to discussion and interpretation of the pattern of significant effects in our attempts to understand the complexity of the findings. In order to summarize the six paired comparisons within the 10 significant outcome variables, a count of significant effects within the ADHD and non-ADHD symptoms domains is informative.

Within the four significant outcome measures in the ADHD symptom domain (parent and teacher ratings of inattention and hyperactivity/impulsivity symptoms), four of the six pairwise

comparisons (MedMgt – Beh, Comb > Beh, CC ≤ Comb, and CC ≤ MedMgt) were significant for almost all (either three or four) of these four measures. Yet for the other two comparisons (Comb ≥ MedMgt or CC < Beh), pairwise comparisons were not significant for any of these four outcome measures. This pattern led to controversy regarding the necessity of behavioral treatment for ADHD symptoms. Indeed, the lack of significance for these two contrasts—Comb ≥ MedMgt (multimodal treatment vs. medication alone) and the CC ≤ Beh (treatment as usual vs. intensive, muticomponent behavioral treatment)— was interpreted to mean that addition of behavioral treatment to medication treatment was not better than medication alone, and that intensive, multicomponent behavioral treatment was not better than treatment as usual in the community (which involved medication in about two thirds of the MTA cases assigned to CC).

For the five domains of outcome beyond ADHD symptoms per se, another systematic pattern was obtained: All of the significant paired comparisons involved the Comb treatment. Comb was better than CC for 5 of the 10 outcome measures (oppositional/aggressive parent and teacher ratings, internalizing parent ratings, social skills teacher ratings, and parent-child power assertion), and Comb was better than Beh for 2 of the 10 outcome measures (oppositional/aggression and internalizing parent ratings). Interestingly, except for parent-rated ODD/aggression, MedMgt was not superior to CC or Beh on any of these outcome measures from the non-ADHD domains, whereas Comb was better than CC on most of the outcome measures and better than Beh on some of them. (It is relevant to note that this pattern was not unique to the non-ADHD domains: as described above, Comb was better than CC on all four of the measures from the ADHD domain and Comb was better than Beh on three of the four ADHD measures, and these effects were larger than for any non-ADHD domain.) In addition, in a detailed analysis of objectively coded parent-child interactions using trained, blind observers, Comb was better than MedMgt (and CC) on positive constructive parenting (Wells et al., 2006), which provided additional evidence of relative benefits from adding behavioral intervention to medication management.

c. Post Hoc Single Composite Outcome Measures

To further consider the hypothesis of multimodal superiority (which was not significant in the main analyses of multiple outcome measures), we explored the use of a single outcome measure in two related but not identical approaches. Swanson et al. (2001) used a narrow composite of just two domains (the ADHD and ODD ratings by parents and teachers). The combination of ratings across symptom domains (inattention, hyperactivity, and ODD) and information sources (parents and teachers) produced a composite measure with about a 40% reduction in variance for each treatment group. For the multimodal comparison, Comb was significantly better than MedMgt ($p \leq .05$) with an effect size of 0.26. Conners et al. (2001) used a broad composite that combined standardized scores for the ADHD and ODD domains plus many others (e.g., internalizing symptoms, social skills, academic performance, and parenting style at home). On this composite, the Comb versus MedMgt comparison was statistically significance ($p < .012$) with an effect size of 0.28.

Thus, in these two articles, we reported that Comb was statistically superior to MedMgt in analyses of a composite score, even though this critical multimodal superiority comparison was not significant for any of the individual outcome measures included in the composites. Two methodological explanations clarify this quirk: (a) as expected from the Spearman-Brown principle, the composite of multiple measures increased the precision of measurement by reducing the treatment group standard deviations for the single outcome measure compared to the separate measures (and thus increased effect size and statistical significance); and (b) the use

of a single outcome measure rather than multiple outcomes required less severe adjustment of significance levels for the statistical tests performed (and thus a smaller *F* or *t* value would exceed the cutoff for statistical significance).

The findings from these two supplementary reports contrast with the findings of primary analyses because they suggest superiority of multimodal treatment whereas the primary analyses did not. These two reports may have generated some confusion despite their consistency, because they offer different reasons for using a composite measure. Swanson et al. (2001) emphasized a statistical reason: The combination of domains and sources increases precision, so that for a single outcome measure the effect size was increased (e.g., from less than 0.16 to 0.26). Conners et al. (2001) suggested that the small effects across non-ADHD outcome measures (all with very small individual effect sizes) combine with the ADHD outcome measures (with the largest individual effect sizes across all measures) to yield significant difference between Comb and MedMgt.

Also, different methods were used for comparing the treatments. Conners et al. (2001) continued with the pairwise comparisons used in the primary analyses (see MTA Cooperative Group, 1999a), whereas Swanson et al. (2001) adopted the orthogonal comparison approach. For the pairwise comparisons, Conners et al. (2001) used the Bonferonni adjustments to the significance level ($p < .05/6 = .0083$), but for the orthogonal comparisons that decompose the overall effect of treatment into nonoverlapping components, Swanson et al. (2001) used the unadjusted significance level ($p \leq .05$). For four treatment conditions, there are six possible pairwise comparisons, but only three independent comparisons (some of which may contrast multiple treatments) can be made. The three comparisons were proposed by Swanson et al. (2001) for exploratory analyses to make the following contrasts: (a) the MTA Medication Algorithm contrast by comparing the average of two groups with this method of intensive medication management as a component of assigned treatment to the average of the other two groups [(Comb + MedMgt) – (Beh + CC)]; (b) the Multimodal Superiority contrast that was the same as one of the pairwise comparisons (Comb > MedMgt); and (c) the Psychosocial Substitution contrast by comparing the generally unavailable intensive behavioral treatment in relation to generally available community standard treatment that in most cases included stimulant medication (Beh – CC).

d. Evaluation of Non-ADHD Outcome Measures

In the primary article (MTA Cooperative Group, 1999a), exploratory secondary analyses suggested that multimodal treatment may be superior to medication alone in some subgroups, but with our specific tests of hypotheses in our design we failed to detect the difference. On the multiple statistical tests performed, comparisons of both Comb and MedMgt to the CC condition produced different patterns of significance across the outcome domains not reflecting ADHD symptoms: Comb but not MedMgt was superior to CC for reading achievement, parent-rated oppositional- defiant behavior, parent-rated internalizing symptoms, teacher-rated social skills, parent-child relations, and reduction in comorbidity. On an objective measure of parenting, Comb was superior to MedMgt. These exploratory analyses suggest that in the MTA design, if we had emphasized outcome measures that go beyond assessment of ADHD symptoms we may have revealed additional evidence in favor of multimodal superiority. Even though the difference was larger for ADHD symptoms than for any of the non-ADHD symptom measures, the analyses of patterns of effects suggests that the lack of significant differences for comparisons of MedMgt versus CC was more frequent in the non-ADHD outcome measures compared to the measures based on ADHD symptoms.

e. Acceptance of the Null Hypothesis

It is important to note that acceptance of null hypotheses is not recommended. We believe that much of the confusion about the findings from the MTA comes from the incorrect interpretation of non-significant results as proofs of the null hypothesis of equivalence. As our statistical consultants warned, "The absence of evidence should not be taken as evidence of absence." Instead, in addition to the test of a priori hypotheses (e.g., in the end-of-treatment analyses, the planned tests of treatment × time interactions followed by the six paired-comparisons of the four groups), post hoc exploratory analyses are recommended to consider possible methodological weaknesses in current research and to direct improvements for future research (Kraemer et al., 2002). For example, we used moderator analyses to address the issue of relative benefit of Comb over MedMgt (the multimodal superiority effect) and Beh over CC (the psychosocial substitution effect). In the secondary article (MTA Cooperative Group, 1999b), we evaluated outcomes in subgroups with and without comorbid anxiety disorder, and in the subgroup with comorbid anxiety, Beh was superior to CC and was not inferior to MedMgt. This supported our decision not to accept the null hypothesis that intensive Beh treatment of the MTA was no different than treatment as usual, which could lead us or others to discard an effective treatment that might be effective. Instead, as recommended by Kraemer et al. (2002), we calculated effect size and interpreted the estimated magnitude of the observed effects in secondary analyses, which suggested some stratification factors that might be considered in a future RCT.

In all of these analyses in the primary, secondary, and additional articles, the effect size for multimodal superiority—the magnitude of the relative superiority of Comb over MedMgt—was not large (i.e., from 0.25 to 0.28). However, this estimated magnitude of effect suggests a new hypothesis that could be tested in a future study. For example, the pattern of significant effects in the multiple comparisons (described above) and the estimates of effect sizes suggest that the MedMgt falls between the extremes of Comb and CC but closer to Comb than to CC. With a different design and greater statistical power, the comparisons of the MedMgt group to one or both of these extreme groups might be statistically significant.

The secondary and additional articles about the effects at the end of the treatment phase of the MTA all address post hoc hypotheses that are interesting, and together the findings may even seem compelling, but they do not have the same standing as the findings based on the a priori hypotheses that were specified by the design of the MTA. By the rules for evidence-based medicine, the positive (i.e., statistically significant) results of the primary ITT analyses of a RCT should take precedence. Unfortunately, in some instances, interpretations of interesting but complex findings from the many exploratory analyses of the MTA may have resulted in confusion and controversy rather than (as intended) clarification of the primary findings of the MTA.

f. Commentaries on the End of Treatment Findings

The initial commentaries and critiques were based on opinions about design characteristics of the MTA. The commentary that accompanied the initial reports (Taylor, 1999) addressed the usual two areas by considering strength (large sample, good design, successful implementation, and sophisticated analyses) and limitation (the definition of long-term, the nature of treatment-as-usual, and tests of null hypotheses). In addition this commentary addressed the interesting issue of dependence on cultural context for definition of treatment-as-usual (which may be similar to MedMgt in the United States but similar to Beh in the United

Kingdom) and for multimodal superiority (which may be defined by Comb vs. MedMgt in the United States but Comb vs. Beh in the United Kingdom). We agreed with all of these percipient comments.

Commentary by Boyle (1999) focused on limitations that we considered to be strengths. For example, he criticized the use of randomization on the grounds that it may have restricted entry into the study and rendered the sample atypical of clinical practice) and high-intensity treatments (which were not generally available and thus lacked ecological validity).

Commentary by Jensen (1999) explained the rationale behind the use of the RCT methods to conduct an efficacy study of intensive interventions rather than an effectiveness study of available interventions, and the feasibility of implementing components of the MTA algorithms to improve treatment-as-usual. He also focused on misrepresentations and mischaracterizations of the MTA, including the claims that the Beh component was ineffective and that the Comb treatment was not superior to MedMgt.

Commentary by Cunningham (1999) focused on the viability of the behavior therapy used in the MTA, the relative benefits of it compared to pharmacological treatment, and the possible advantages of combined treatment. He pointed out that secondary analyses will be required to address these questions in an adequate fashion. We agreed with these points, and since then we have provided many secondary analyses.

Commentary by Pelham (1999) focused on aspects of the design that may have predisposed the MTA in favor of pharmacological treatment (which reached highest intensity at the end of the 14-month treatment phase due to medication adjustments) over psychosocial treatment (which was faded at the end of the 14-month treatment phase).

Commentary by Barkley (2000) focused on our lack of use of theory of the underlying deficit in ADHD (with an example of his own theory of "inhibition deficit"). He speculated this was the primary reason that "the Beh treatment was ineffective." In response to this commentary, Swanson et al. (2002) pointed out this interpretation was based on acceptance of the null hypothesis and confusion about the comparisons of outcomes over time in the regression analysis framework, which showed significant effects of time for all groups but relative differences in degree of improvement across the treatments.

Breggin (2001) focused on the lack of a placebo control and weak effects based on blind observers and concluded that the evidence from the MTA did not justify a recommendation of medication as the "first line" treatment. This was followed by commentary by Klein (2001), who concluded that the evidence from the MTA did not justify a recommendation of multimodal treatment by adding Beh to MedMgt. In a response to this commentary, Jensen (2001) pointed out that these conclusions—based on the same findings—were polar opposites.

Greene and Ablon (2001) focused on ecological validity of the MTA treatments, suggesting that the lack of tailoring of the behavioral treatments may have underestimated the effects of this modality in clinical settings. In response to this commentary, Wells (2001) addressed the problems involved in "tailoring" and the lack of effectiveness of tailoring treatments in other large trials., and Hoza (2001) related the treatments and outcomes of the MTA to prior studies of social, family and motivational processes.

g. The Main Confusions and Controversies at the End of Treatment

Why were the primary findings reported by the MTA Cooperative Group (1999a, 1999b) controversial? Based on the evidence at the end-of-treatment, the multimodal treatment did not provide significantly greater benefits than unimodal treatment with stimulant medication for any of the outcome measures. This pattern of outcome suggested that many children treated with the

medication as prescribed by the MTA algorithm would not require the intensive behavioral interventions of the MTA to achieve the full effect of treatment. The assertion that intensive behavior therapy was ineffective was controversial.

Some of the secondary and additional findings were intended to dispel confusion and resolve some controversies about the primary findings about the ineffectiveness of the Beh treatment, a conclusion that was based on the premature acceptance of the null hypothesis. We did not provide evidence of effectiveness of Beh treatment in general (i.e., in the primary analyses of the MTA), but moderator analyses suggested that for subgroups with certain characteristics Beh was more effective than for other subgroups. For example, analyses of moderators of treatment effects suggested that in the subgroup of children with ADHD and comorbid anxiety, the Beh treatment appeared to be particularly effective (and not ineffective as erroneously interpreted by some from the primary findings). Furthermore, when non-ADHD outcomes are considered, in socioeconomically disadvantaged families the Comb treatment was superior to MedMgt with respect to social skills (not equal as suggested in the primary analyses). These secondary analyses suggest that enhanced response to the Beh treatment may have occurred in some subgroups of participants. We did not focus our secondary analyses only to explore effects of the MTA behavioral algorithm. We also used secondary analyses to explore effects of the MTA medication algorithm. Mediator analyses suggested whether the participants accepted the assignment to treatments with medication or followed the MTA medication algorithm "as intended" affected long-term outcome, and analysis of the CC group suggested that the way community clinicians prescribed and monitor medication as also affects long-term outcome in this comparison group.

Taken together, these supplementary analyses at the end of the 14-month treatment phase suggested that long-term relative effects of the different modalities of treatments may depend on who was treated (e.g., the characteristics of the subgroup—those with anxiety may have enhanced response to the Beh treatment), how they are treated (e.g., the nature of the pharmacologic protocol— those with more intensive monitoring and adjusting of dose may have enhanced response to treatment with stimulant medication), and what outcome measures were evaluated. Nonetheless, the implication that the Beh treatment was not a necessary component of treatment of ADHD was one of the most controversial aspects of the initial reports of the end-of-treatment findings (MTA Cooperative Group, 1999a).

C. Publications After the First Follow-Up (MTA Cooperative Group, 2004a, 2004b)

1. Evidence: ITT Analyses

At the first follow-up assessment, 10 months after the end of treatment by MTA protocols, recommendations were made and participants were left to obtain their own individualized treatment regimens. Despite this, ITT analyses were used again to evaluate the persistence of effects of the randomly assigned treatments (MTA Cooperative Group, 2004a). However, some changes in the analysis framework were made.

First, while the initial definition of long-term in the MTA was just beyond a year (14 months), the revised definition for longer term was now 2 years (24 months). Thus, in these analyses the evaluation of long-term focused on outcome 10 months after the end of our intensive treatment phase when interventions were delivered by the MTA protocols. Rather than evaluate regression equations for each group and compare slopes by testing the treatment × time interaction, analyses of the 24-month outcomes (with the baseline observations used as

covariates) were performed to evaluate the persisting effects of assigned treatment at this follow-up assessment.

Second, a different method for making multiple comparisons was adopted. Benefiting from our previous experience, we switched from the pairwise comparison method to the orthogonal comparison method. The 3 orthogonal comparisons developed by Swanson et al. (2001) were adopted: (a) the MTA Medication Algorithm contrast, which compares the average of two groups assigned to this method for intensive medication management to the average of the other two groups [(Comb + MedMgt) – (Beh + CC)]; (b) the Multimodal Superiority contrast, which is the same as one of the pairwise comparisons (Comb > MedMgt); and (c) the Psychosocial Substitution contrast, that compares the generally unavailable intensive behavioral treatment to the generally available community standard treatment that in most cases included stimulant medication (Beh – CC).

Third, in these follow-up analyses we changed the number of outcome measures from 14 to 5, which imposed less stringent per-comparison alpha correction procedures. Instead of treating each source (parent and teacher) as a separate measure on ratings from the ADHD, ODD, and social skills domains, we treated source (rater) as a factor in the analysis. If the main effect of factor was not significant, then collapsing across the two levels allowed us to combine (averaged) the parallel ratings from the two sources and reduce the number of outcome measures. Also, based on the principle of parsimony, we refined the outcome domains, using 2 domains that related to symptoms of psychiatric disorder (ADHD and ODD symptom ratings) and 3 domains related to functional impairments (Social Skills, Parental Discipline, and Reading Achievement). Together, these changes allowed for less stringent adjustment for multiple comparisons ($p \leq .05/5$ or $p \leq .01$). We first evaluated the overall effect of treatment on 24-month outcome, and if this was statistically significant at $p \leq .01$, then we performed the 3 orthogonal comparisons.

In the primary analyses, the main effect of assigned treatments was significant for the ADHD ($p = .0001$) and ODD ($p = .0019$) domain measures, but not for the outcome measures of Social Skills ($p = .06$), Discipline ($p = .05$), or Reading ($p = .38$). The subsequent orthogonal comparisons for the two significant domains revealed that only the MTA Medication Algorithm effect (Comb + MedMgt – Beh + CC) was significant. To evaluate the magnitude of this significant effect at the 24-month follow-up, the effect size was contrasted with the effect size at the 14-month assessment. This showed that the magnitude of the relative superiority of the MTA medication algorithm was reduced by 50% from a medium effect size in the end-of-treatment analyses (0.6) to a small effect size in first follow-up analyses (0.3).

We observed that between the 14-month and 24-month assessments, the MedMgt and Comb groups showed a slight increasing trend in symptom severity, whereas the Beh and CC groups remained constant. We performed mediator analyses to test whether this loss of relative advantage of the MTA Medication Algorithm could be attributed to current medication use. To accomplish this, we used the report of current medication use at the 24-month assessment point as a covariate. (Two definitions of *current*—anytime during the 10-month follow-up or during the 30 days before the 24-month assessment—were considered, but both produced the same effect.) The covariate was significant for the analysis of ADHD symptoms, indicating that current medication improved outcome at the 24-month assessment. However, even after adjustment for current medication use, the overall treatment effect and the MTA Medication Algorithm contrast remained highly significant. These findings suggest that after treatment was no longer provided or monitored by the MTA staff, choices about taking medication during the

follow-up explained part but not all of persisting advantage of the MTA medication algorithm (MTA Cooperative Group, 2004a).

2. Interpretation: Naturalistic Subgroup and Physical Growth Analyses

a. Naturalistic Subgroups Based on Actual Treatment

The secondary analyses of the first follow-up (MTA Cooperative Group, 2004b) extended the primary analyses by (a) evaluating patterns of actual treatment over time in "naturalistic subgroups" (rather than assigned treatment in the primary analyses) and (b) evaluating change in outcome between the 14-month and 24-month assessment points (rather than status at the 24-month assessment point as in the analyses reported in the primary article).

Naturalistic subgroups were formed by considering medication status (Med or NoMed) at the end-of-treatment and first follow-up assessment points. We identified four subgroups defined by a pattern of consistent use (Med/Med) or nonuse (NoMed/NoMed) of medication at the 14-month and 24-month assessment points, or a pattern of stopping (Med/NoMed) or starting medication (NoMed/Med) during the 10-month follow-up. In this initial method to establish naturalistic subgroups, we did not consider medication status of the participants before entering the MTA. (Later we realized that this may be important and attempted to evaluate this factor [see Swanson et al., 2007b].) Our operational rule for establishing medication status was whether medication was taken or not during the 30 days prior to the assessment point. Because all participants had a 30-day "wash-out" period without medication before the baseline assessment, we considered all participants to have the same status (NoMed) at the baseline assessment, even though about one third of the participants had a prior history of medication use.

In short, these secondary ITT analyses suggested that the Comb and MedMgt groups manifested greater change than the Beh and CC groups, since they showed greater deterioration during the follow-up from the 14- to the 24-month assessment point. In these secondary analyses, we then added naturalistic subgroup membership as a mediator. After this adjustment, which accounted for the improvement in symptoms of ADHD in the subgroup that started medication and the deterioration in the subgroup that stopped medication, the effect of assigned treatment was no longer significant. In combination with the differential adherence to assigned treatments (i.e., more cases in the MedMgt and Comb groups stopped medication and more cases in the Beh and CC groups started medication), this implicated choice about medication use during the 10-month follow-up as a factor that contributed to the 50% loss of relative superiority of the MedMgt and Comb over the Beh and CC groups (MTA Cooperative Group, 2004b).

The different methods used for the primary analyses (absolute scores at the 24-month assessment point, with current medication as a mediating variable) and the secondary analyses (change scores from the 14-month to the 24-month assessment points, with naturalistic subgroup based on pattern of medication use as a mediator) may be the source of some confusion about the findings of the MTA at the first follow-up. However, the findings based on change scores in the secondary article (MTA Cooperative Group, 2004b) converged with the findings based on absolute scores in primary article (MTA Cooperative Group, 2004a). Both suggested that the persisting but reduced advantage (which also could be called the partial loss of relative superiority) of the MTA Medication Algorithm was mediated by the pattern of medication use over time. Furthermore, both concluded that the subgroup being treated with medication at the follow-up assessment point (i.e., at the time when the ratings of ADHD symptoms were provided) had better outcome than the subgroup not being treated.

b. Evaluation of Stimulant-Related Growth Suppression

One of the most controversial findings at the first follow-up was from a post hoc analysis in the secondary article (MTA Cooperative Group, 2004b), which suggested a possible long-term side-effect of treatment with stimulant medication on physical growth. Because the hypothesis of stimulant-related growth suppression suggested by Safer et al. (1976) had been largely discounted in the years since its original publications (Roche et al., 1979; Spencer et al., 1996; see NIH Consensus Conference, 1998), it was not an a priori hypothesis and thus was not specified as a primary question to be addressed by the MTA or even evaluated in the initial analyses and main report (MTA Cooperative Group, 1999a, 1999b). Thus, post hoc analyses were performed to evaluate physical size at the end-of-treatment. Because age and sex were balanced across the assigned treatment groups, we used absolute measures of size (cm and kg) rather than relative measures of size (*z* scores) in these analyses. This might be considered a weakness by some (see Spencer et al., 1996), but others (see Kraemer et al., 2000) recommend the use of raw scores not standardized scores in some situations, such as when the norms used for standardization are cross-sectional. Even though growth norms are based on cross-sectional data, in the next report on growth in the MTA we used relative measures of size (see Swanson, Elliott, et al., 2007).

The same strategy that we applied for the ADHD outcome measures (see above) was used for the analysis of height and weight (i.e., ITT analysis of assigned treatment on change in size from baseline to the 14-month end of treatment assessment point). These analyses revealed that for the groups assigned to treatments that included the MTA medication algorithm, the average gain in height (Comb = 4.85 cm and MedMgt = 4.75 cm) was less than for the other two groups (Beh = 6.19 cm and CC = 5.68 cm). The same pattern held for gain in weight (Comb = 2.52 kg and MedMgt = 1.64 kg vs. Beh = 4.53 mg and CC = 3.13 kg). Thus, the "gold standard" for evidence-based medicine—ITT analysis of an RCT—provided clear evidence of stimulant-related growth suppression. For the most informative comparison of the two unimodal treatments (MedMgt with assignment of medication and Beh without assignment of medication), we observed a reduction of growth velocity by about -1.23 cm/yr in height and -2.48 kg/yr in weight at 14 months. These might be considered conservative estimates, because there was imperfect acceptance and compliance with assigned treatments, as noted above (see Marcus & Gibbons, 2001).

By the end of the 10-month follow-up, ITT analyses no longer showed a stimulant-related growth suppression effect based on a comparison of the assigned groups or the MTA medication algorithm comparison (MTA Cooperative Group, 2004b). However, the expected growth rebound was not detected by the analysis of 14-month to 24-month change for the assigned treatment groups: All assigned groups had about the same gain in height and weight (Comb = 5.69 cm and 5.28 kg; MedMgt = 5.69 cm and 5.06 kg; Beh = 6.16 cm and 4.98 kg; CC = 5.79 cm and 4.58 kg). On the other hand, analysis of actual treatment defined by the naturalistic subgroups revealed that a subgroup of consistently treated children (i.e., those on medication at both the 14-month and 24-month assessments) relative to an untreated subgroup (i.e., those who were not treated with medication at the 14-month or 24-month assessments) showed reduced gain in height and weight, which suggested additional stimulant-related growth suppression during the first 10-month follow-up period. In the naturalistic subgroups, the apparent effects of actual treatment on reduction in annual growth velocity (12-month estimates from the initial 14-month and subsequent 10-month intervals) was similar for the height reduction documented during the treatment phase (-0.9 cm/yr) and follow-up phase (-1.02 cm/yr), but the apparent

reduction for weight was less in the follow-up phase (-1.22 kg/yr) than documented in the treatment phase (-2.55 kg/yr).

3. Qualification: Loss of Effectiveness or Lack of Maintenance

a. Partial Loss of Relative Benefits of Medication

We speculated that the partial loss of relative superiority may have been related to the continued use of ITT analyses of the randomized groups even after the assigned treatments were no longer provided. In fact, these analyses evaluated the effects of assigned treatments even though at the end of the treatment phase we recommended for many cases that the components of assigned treatment be changed (i.e., at the 14-month assessment we recommended adding medication in 74% of the cases in the Beh group). Also, we speculated that our methods were based on relative comparisons of conditions superimposed on a much larger effect of general overall improvement over time (i.e., a dramatic reduction in severity of ADHD symptoms), making it more difficult to detect positive effects of medication. This speculation was supported by supplemental analyses of actual treatment that added a covariate for interim medication use, which revealed a significant positive effect of medication on outcome at the 24-month assessment—in addition to the smaller but still significant effect of assigned treatment.

Nonetheless, the main controversy at the first follow-up was about lack of maintenance of medication. In the MedMgt and Comb groups, the number of cases continuing medication decreased, and we speculated that deterioration in those stopping medication might account for some of the apparent loss of effectiveness of these conditions compared to the Beh condition, which showed an increase in the number of cases treated with medication. Thus, one aspect of this controversy was about changes in actual treatment, which we speculated may operate to discredit an effective treatment. We predicted that if the trend of between-group differences in starting and stopping medication continued, then the assigned groups would converge in later assessments of outcome. This prediction was tested in the evaluation of outcome in the next follow-up (see Jensen et al., 2007).

b. The Main Confusions and Controversies at the First Follow-Up

Why were these primary findings of the first follow-up controversial? Some children (particularly those in the MedMgt and Comb groups) deteriorated during follow-up when the assigned treatments were no longer provided by research staff according to the MTA protocols. This was reflected in ITT analyses of randomly assigned treatment groups, which suggested that the relative benefits of medication were waning over time. This pattern was different than expected from other long-term follow-up studies (e.g., see Abikoff et al., 2004), which showed persistence of benefits across 2 years of treatment when intensive medication was provided by research staff as part of the study's protocol.

Also, ITT analyses and secondary analyses of actual treatment appeared to contradict some well-established and long-held beliefs about physical size and growth of children with ADHD. For example, the MTA data on physical size did not support the generally accepted theory of disorder-related rather than treatment-related differences in physical size that hypothesized children with ADHD had a "slow tempo" or delay in growth (Spencer et al., 1996). In addition, the persistence of growth suppression that appeared to accumulate over 2 years did not support the consensus opinion of two blue ribbon panels, which had concluded that long-term stimulant-related growth suppression was not of significant concern (NIH Consensus Conference, 1998; Roche et al., 1979).

Thus, the analyses at the first follow-up suggested that the balance of benefits and side-effects was different that earlier reported—with smaller benefits and greater side-effects than expected based on prior findings and the literature.

D. Second Follow-Up (Jensen et al., 2007; Molina et al., 2007; Swanson, Elliott, et al., 2007; Swanson, Hinshaw, et al., 2007)

1. Evidence: ITT and Moderator Analyses

The analyses of the first follow-up 24 months after randomization (see above) suggested that the apparent reduction of the relative effects of stimulant medication may reflect lack of maintaining an effective intervention more than reduction in the effects of medication. Even though we predicted the convergence of assigned groups by the 36-month assessment (MTA Group, 2004b), we expected to account for this by the patterns of long-term maintenance of treatment with medication in our analyses of outcomes over this additional time interval.

As discussed repeatedly above, in the RCT design ITT analyses of assigned treatment are evaluated over the long-term even when treatments are no longer provided or have changed. In the primary article about the 36- month outcomes, this approach was used again to evaluate differences across the assigned treatment groups at the 36-month assessment. Benefiting from experience gained from the end-of-treatment and the first follow-up evaluations, we used four of the same outcome domains described previously (parent-teacher ratings of ADHD and ODD symptoms; social skills rating; reading achievement), plus one additional outcome measure that was considered appropriate for the older age of the participants (the Columbia Impairment Scale, which replaced the Parental Discipline measure used in the analysis of the first follow-up). Also, we used the same three orthogonal comparisons adopted for the 24-month assessment to test for persisting relative superiority of the MTA medication algorithm (Comb +̣ MedMgt >̣ Beh̩ + CC), as well as for Multimodal Superiority (Comb >̣ MedMgt; Comb >̣ Beh) and for Psychosocial Substitution (Beh >̣ CC).

The ITT analyses show that the assigned groups did not differ significantly on any of the five outcome measures (Jensen et al., 2007). The comparison of effect size indicated that the large relative superiority of the MTA medication algorithm at the 14-month assessment was negligible at the 36-month assessment. That is, despite the large superiority of the medication algorithm at 14 months, and its continued presence (though reduced size) at the 24-month assessment, by the 36-month assessment there was essentially no difference between those participants who were and were not initially randomly assigned to a treatment that included intensive medication management according to the MTA protocol. However, it is important to note that this effect was superimposed on a very large overall improvement for all of the assigned treatment groups (e.g., for the baseline-to-36-month change, the effect sizes estimates were from 1.6 to 1.7 across the four treatment groups). Thus, as expected from most longitudinal studies in the literature, in the MTA follow-up the severity of ADHD symptoms decreased over time, but this was independent of assigned treatment.

These findings at the 36-month assessment confirmed the prediction stated in the prior report (see MTA Cooperative Group, 2004b): The residual relative superiority of the MedMgt and Comb treatments over the Beh and CC treatments had completely dissipated by 22 months after the end of the treatment phase (Jensen et al., 2007).

In the primary article, mediator analyses were applied to evaluate possible reasons for the complete loss of relative superiority of the MTA medication algorithm (Jensen et al., 2007). From the SCAPI, the reports of actual treatment received revealed that there was a 24-36 month

increase in the percentage of participants using medication in the Beh group (from 34.9% to 45.2%), but this percentage in the other treatment groups remained nearly constant (see Jensen et al., 2007, Table 2: 71.0% to 70.4% for Comb; 71.6% to 71.8% for MedMgt; 62.3% to 62.4% for CC: the last dose equivalents listed in the table are averages that include 0 doses for participants not treated with medication at the 36-month assessment). Analysis of covariance for ADHD symptom severity revealed that actual medication use (percentage of days treated) was significant as a main effect and interacted with time. However, the direction of this effect was different for the two follow-up intervals: Greater medication use produced better outcome at the 14-month and 24-month assessment points, but at the 36-month assessment point there was no significant mediating effect of medication use for the overall baseline to 36-month follow-up. Thus, our prediction that loss of relative superiority would be related to maintenance of treatment was not confirmed. Intriguingly, a post hoc analysis of 24-36 month change in symptom severity revealed that medication use over this follow-up interval was related to deterioration (*increase* in symptom severity) rather than benefit.

Also, in the primary article (Jensen et al., 2007) we performed moderator analyses to evaluate comorbid anxiety and other subgrouping factors present at baseline as moderators of treatment response, but these did not reveal subgroup differences. Thus, the initial enhanced response to the Beh treatment in children with comorbid anxiety (see MTA Cooperative Group, 1999b) was not observed in the long-term follow-up.

2. Interpretation: Multiple Secondary Articles

a. Departure From Prior Pattern

In a departure from our prior strategy of publishing two main papers under the group name, we decided that multiple secondary articles were necessary to address possible explanation for these findings and to address important additional outcome measures as the sample approached adolescence, which required considerable time and effort of working groups that should be recognized by naming authors. Also, when using standard methods to search for articles published on the MTA, the typical strategies (i.e., author's names) do not always locate or identify the published articles, making a review of the complete literature difficult.

b. Selection Bias and Individual Differences

In the primary article by Jensen et al. (2007), exploratory analyses were reported as well as the main ITT analyses, which suggested that a small subgroup who had stopped medication between 24 and 36 months were doing better at 24 and 36 months than another small subgroup who started medication between 24 and 36 months. This seemed to indicate that self-selection for medication might explain the apparent lack of relative superiority at 36 months. However, these subgroups were small. The results were reported electronically in the Article Plus website, and a reference was made to more detailed analyses in the companion article by Swanson, Elliott, et al. (2007).

In the initial secondary article (Swanson, Hinshaw, et al., 2007) we addressed in more depth the hypothesis that selection biases might account for the loss of relative superiority of medication. If the most severe cases, expected to have poorer outcome than the less severe cases, were preferentially treated with medication at the 36-month assessment, then this factor may account for its apparent loss of efficacy. The statistical method selected to evaluate this hypothesis was the propensity score analysis. This propensity analysis relies on the assumption that selection biases can be modeled as a simple linear combination of multiple, complex

variables, such as ethnicity, previous experience with medication, treatment response, SES, and other variables, which may or may not be correct. As described and discussed in Swanson, Hinshaw, et al. (2007), a propensity score was created for each participant based on variables associated with the tendency to take medication. We used this procedure to test the hypothesis that medication use might be associated with both factors—the propensity to take medication and to have poor outcome. If present, this association could result in a bias when comparing subgroups that were treated and untreated with medication, and could theoretically mask any relative superiority of treatment with medication.

However, counter to our hypothesis, the severity of ADHD symptoms was *not* related to the use of medication at the 36-month assessment: The quintile representing those cases most likely to use medication at the 36-month assessment had lower rather than higher ratings of ADHD symptoms at the 14 and 24 month assessment points. Also, when subgroups of cases with and without actual use of medication were compared across five levels of the propensity score (quintiles), there was no evidence of the superiority of actual use of medication in any quintile (Swanson, Hinshaw, et al., 2007). Thus, the propensity score analyses did not provide statistical support for the basic hypothesis that selection bias explained the lack of a significant medication effect in the MTA follow-up.

Additional analyses of change over time were performed using a statistical procedure called growth mixture model (GMM) analysis (see Hedeker & Gibbons, 2006). In this context, the use of the word *growth* does not refer to physical growth but to the trajectory of change over time in any outcome measure. As described and discussed in Swanson, Hinshaw, et al. (2007), the GMM approach was applied to determine if the trajectory of ADHD symptom severity over time was homogeneous across all participants in the MTA follow-up or whether more than one trajectory provided a better explanation for change over time. The best fitting model was for three latent classes, each with a different pattern of change in ADHD symptoms over time. Class 1 (34% of the sample) showed an trend of gradual improvement (decrease in ratings of ADHD symptom severity) that was small at the 14-month assessment but continued to the 36-month assessment. Class 2 (52% of the sample) showed a large improvement at the 14-month assessment that was maintained through the 36-month assessment. Class 3 (14% of the sample) showed a large initial improvement followed by a trend of deterioration over time that resulted in a return to the baseline level of ADHD symptom severity by the 36-month assessment. This is consistent with common sense and clinical experience, as it indicates that not all children with ADHD are the same and that individual differences in treatment response and development should be considered.

Analyses were performed to examine the possibility of different effects of assigned and actual medication use in these three classes. Even though the overall comparisons of subgroups with and without consistent medication revealed no overall evidence for the relative superiority of medication, the multiple classes from the growth mixture model analysis allowed for within-class comparisons to test the hypothesis that counteracting effects across the classes might mask the relative superiority of medication within certain classes.

The medication effect was evaluated within each of these latent classes by comparing the class members with consistent high use of medication over time to the class members with consistent low use of medication. All three latent classes showed a significant relative superiority of consistent high use over low use at the 14-month assessment, but for Classes 2 and 3 this effect dissipated and was absent by the 36-month assessment (i.e., the same pattern reported by Jensen et al., 2007). However, for Class 1, the relative superiority of medication (defined by

consistent high vs. low medication use) was significant at the 14-month assessment ($p = .041$), and it became larger and was still significant at $p = .001$ at the 36-month assessment. This pattern suggests not only maintenance of the medication effect but an increase in its magnitude over time. Because this pattern was present in a minority of the cases (34% of the sample in this class), it was masked in the primary analyses by the opposite pattern in the majority of cases (i.e., a large initial relative superiority of medication that dissipated over time in the 66% in the two classes comprising a majority of the sample). The analysis of the baseline measures showed that Class 2 differed from both Class 1 and 3 but there was no clear distinction between 1 and 3. The participants in these two classes had patterns of high scores that indicated much greater adversity on demographic measures and severity on symptom measures than the participants in Class 2.

Also, the most favorable trajectory (shown by Class 2) had a preponderance of children originally assigned to treatment with the MTA medication algorithm (62% of the Comb and 55% of the MedMgt, for a total of $n = 169$, vs. 46% of the Beh and 45% of the CC, for a total of $n = 129$), which suggested a possible residual benefit in some cases even when actual medication use had stopped.

c. Stimulant-Related Growth Suppression

The third article focused on the hypothesis of stimulant-related growth suppression (Swanson, Elliott, et al., 2007). Based on the surprising finding of stimulant-related growth suppression uncovered by post hoc analyses in the second main report (MTA Cooperative Group, 2004b), we continued our analyses of physical growth. For the analyses of the data through the 36-month assessment point, we improved our methods. In emerging naturalistic subgroups based on treatment choice at each successive assessment point, we identified individuals who were never, newly, consistently, or inconsistently treated with medication. We performed more sophisticated analyses of height and weight by using standard scores based on national growth norms (*z* height and *z* weight) as the outcome measures, rather than absolute height and weight as in the prior analysis.

In the regression analyses, the interaction of naturalistic subgroup with time was significant for both height and weight. This provided the statistical support for the hypothesis that the pattern of treatment with stimulant medication was associated with the growth velocity over time. We reported a surprising pattern of growth in the participants never treated with medication (i.e., an untreated clinical control group). Compared to population norms, this group was physically much larger than expected from U.S. norms (or by comparison to classmates in our own local normative comparison group consisting of classmates of the ADHD participants). Also, this never-treated subgroup showed growth acceleration (an increase in *z* scores) over time. In contrast, the newly treated subgroup (initially stimulant-naïve but then consistently treated after entering the MTA) showed growth suppression (a decrease in *z* score). The magnitude of the difference in growth between these two diverging groups over the 36-month follow-up was about 2 cm, which occurred during the first 24 months and then stabilized, with no recovery (i.e., no growth rebound) but with no additional growth suppression either. Whether stimulant-related growth suppression is permanent or will be made up at the end of the growth cycle is a question to be answered by future analyses of data from the 12-year assessment when the participants will be between 19 and 21 years of age.

In the formation of naturalistic subgroups at this point in the follow-up, we took into consideration prior history of medication to form the naturalistic subgroups. Even though this was very limited information (i.e., report of use of medication within 30 days prior to the

baseline assessment), our analyses revealed an interesting finding: The subgroup treated with medication prior to entry into the MTA and then consistently treated with medication in the MTA protocol were physically much smaller at baseline than the other naturalistic subgroups. It is possible that prior treatment with medication resulted in this difference. However, these naturalistic subgroups were not established by randomization, so it is also possible that a selection bias was present, and the smallest ADHD participants in the MTA were selectively treated with medication at an early age.

d. Delinquency and Substance Use

The fourth article focused on a new set of outcome measures (Molina et al., 2007). With adolescence impending by the 36-month follow-up, we analyzed severity of delinquent behavior and substance use (Molina et al., 2007). Delinquent behavior was coded along a 5-point continuum of severity using several parent- and child-reported measures. Twenty-seven percent of the children had engaged in moderate-to-serious delinquent behaviors by the 36-month follow-up, and these rates were not affected by original randomized treatment group assignment. They were significantly higher than the mostly non-ADHD classmate comparison group called the Local Normative Comparison Group ($p \leq .0001$). Not surprisingly, children with higher delinquency scores were more likely to be medicated at follow-up.

For these analyses, substance use was defined as any use of alcohol (more than a sip), tobacco, or marijuana. The observed rates of endorsement were significantly higher among the MTA participants compared to the LNCG participants (classmates of the ADHD children). Counter to the predictions in the literature (see Biederman, Wilens, Mick, Spencer, & Faraone, 1999; Wilens et al., 2003), the participants in the MTA who were treated at a young age with medication did not show any evidence of protection from initial use of substances in our 36-month analyses. Instead, receipt of behavioral therapy was associated with some protection. At the 24-month assessment, the children who received behavioral therapy (Beh or Comb) were less likely to report substance use than the children who did not (MedMgt and CC). This effect was not significant at the 36-month assessment, but the direction of the difference still favored the groups assigned to receive behavioral algorithm of the MTA. These results do not rule out the potential for long-term protective or predisposing associations between treatment and later substance use, abuse, or dependence that emerge at older ages. Thus, further analyses are crucial as the children mature through adolescence into early adulthood when rates of substance use, abuse, and dependence reach their peak.

3. Qualification: Challenging Consensus Views and New Predictions

a. The Main Confusions and Controversies at the Second Follow-Up

The findings of these secondary articles were not consistent with views and expectations about medication effects held by many investigators and clinicians in the field. That is, long-term benefits from consistent treatment were not documented; selection bias did not account for the loss of relative superiority of medication over time; there was no evidence for "catch-up" growth; and early treatment with medication did not protect against later adverse outcomes. It is likely that these challenges to consensus views contributed to confusion and controversy about the long-term outcomes in the MTA.

Why were these primary and secondary findings at the second follow-up controversial? The intent-to-treat analyses suggested that the modest significant advantages we found at the 24-month assessment for the MTA Medication Algorithm were completely lost by the 36-month

assessment. Because this was predicted from the findings at the end of the first follow-up (MTA Group, 2004b), this was not unexpected or controversial. However, counter to our prediction, when actual treatment with medication was evaluated, the subgroup with current treatment (compared to the subgroup without) showed a tendency toward disadvantage rather than the benefit. This may be the primary reason for controversy at this point in time.

However, we found no support for the hypothesis that selection biases were "carrying" this lack of long-term benefit from medication—even though we used a sophisticated method (propensity score analysis). This method had previously uncovered a participation bias during the treatment phase of the MTA, which was found to affect outcome at the 14-month assessment (Marcus & Gibbons, 2001), but when applied to the long-term follow-up phase, this method did not confirm our predictions (see Swanson, Hinshaw, et al., 2007).

Also, based on the literature, we expected that initial growth suppression might be followed by growth rebound, but by the 36-month assessment this pattern had not occurred (Swanson, Elliott, et al., 2007). Still, the initial growth suppression had abated by 36 months, even in consistently medicated youth. Additionally, one strong rationale for the use of stimulant medication was the expectation that this would offer protection against the emergence of substance use or juvenile delinquency, but no evidence to support this hypothesis emerged from the MTA by this point in the follow-up (Molina et al., 2007).

The intention and primary value of research is to discover new knowledge, but when new findings contradict consensus views, confusion and controversy should be expected. This has occurred as we have published the findings of each phase of the MTA so far. It is likely to continue in the future as the follow-up continues and new discoveries are made and disseminated.

b. Subsequent Publications of Multimodal Treatment Studies

Two articles describing effectiveness studies in Europe have addressed some of the critiques and predictions from the MTA efficacy study: Dopfner et al. (2004) and van der Oord et al. (2007).

In a study conducted in Germany, Dopfner et al. (2004) addressed the issues of ecological validity by using an adaptive design for an effectiveness study of treatment strategies similar to clinical practice. A sample of 75 children with diagnoses of ADHD were randomized to behavioral intervention (which was purposely less intensive that the Beh condition of the MTA, which was estimated to be about three times as intensive) or to medication management (which was also purposely less intensive than MedMgt in the MTA, which was estimated to be about two times as intensive). The disposition of the sample was dependent on acceptance of randomization (which was low for assignment to treatment with medication) and response (components were stopped or started dependent on outcomes at four stages of the study). Outcome was evaluated by using change from baseline as an indication of improvement. At the end of treatment, the resulting adaptive multimodal treatment produced a success rate of 55% (compared to 68% for the MTA). Part of the difference was attributed to differences in intensity of treatment components, which was estimated to be about half as intensive as the pharmacological component of the MTA and one third as intensive as the psychosocial component of the MTA. Adjusted for intensity of intervention, these outcomes suggest similarity with one of the main MTA findings (a slight multimodal superiority despite a lower dose of medication), but with a more ecologically valid set of interventions that were less intensive and expensive.

In a study conducted in the Netherlands, van der Oord et al. (2007) addressed the issue of differential intensity of the medication component of the MTA Comb and MedMgt conditions. They evaluated 50 children with ADHD whose medication was titrated using a procedure similar to the MTA double-blind titration trial, and randomly assigned them to receive a 10-week behavior therapy (or not). This defined two groups with multimodal (medication plus behavior therapy) and unimodal (medication only) treatments. The dose of methylphenidate was lower than in the MTA (20.8 mg/day) and the placebo response rate was higher (41%). The intensity of behavior therapy was designed to be less than in the MTA in order to match psychosocial intervention that was practical and available. The outcomes at the end of the 10-week treatment phase showed no difference between the two groups in multiple measures of outcome designed to evaluate symptoms as well as global functioning. The findings from this effectiveness study are consistent with the findings from the MTA efficacy study, and produced estimates of effect size that were comparable. However, the findings from this effectiveness study should be interpreted with the same cautions recommended here for the interpretation of the findings from the MTA, with special caution about acceptance of the null hypothesis.

E. Clinical Relevance

1. Questions and Interim Answers

What do the findings from the MTA offer to clinicians and parents who seek answers to key questions about long-term effects of stimulant medication? Taken together, the primary publications of the MTA provide evidence to answer the following questions.

a. What is the first-line treatment for initiating treatment of children with confirmed diagnoses of ADHD-Combined Type?

Over the 14-month treatment phase of the MTA, intensive management of stimulant medication by the MTA medication algorithm, with or without concurrent behavior modification, provided relative superiority compared to the treatments without this component, at least for the outcome domains related to ADHD and ODD symptom severity. Thus, the evidence from the MTA supports this regime as the first-line treatment (MTA Cooperative Group, 1999a).

b. Does childhood treatment with stimulant medication produce a clinically meaningful reduction in growth?

Stimulant-related growth suppression (about a 20% reduction in annual gain in height) was documented by the gold standard—an intent-to-treat analysis at the end of treatment. Exploratory mediator analyses suggested this effect may accumulate over 2 years to reduce height gain by about 2 cm in children receiving consistent medication, but no further increase in (or rebound from) height suppression occurred in those children remaining on medication for an additional year. However, this may not be a permanent reduction in physical size, as catchup or rebound growth could still be possible in the MTA sample, which was only 10 to 12 years of age at the time of the 36-month assessment.

c. Is the relative superiority of the MTA medication algorithm a long-term effect?

A fundamental purpose of the MTA was to provide evidence about long-term effects. Long-term relative benefits of stimulant medication were documented and appeared to remain constant in the early, middle, and end of the treatment phase of the randomized clinical trial (i.e., at the 3-, 9-, and 14-month assessments). Therefore, when long-term is defined as up to 2 years,

the MTA findings provide evidence of a long-term relative benefit of the initiation of treatment with the MTA medication algorithm (MTA Cooperative Group, 1999a, 2004a). However, if the definition of long-term is extended to 3 years, the relative superiority appears to be lost, so this benefit does not seem to be a permanent and thus could be considered a temporary effect (Jensen et al., 2007; Swanson, Elliott, et al., 2007; Swanson, Hinshaw, et al., 2007).

d. Does the relative superiority last as long as the treatment is maintained?

Despite a relatively high rate of medication use (70%), the relative superiority of having experienced the MTA medication algorithm was negligible and nonsignificant by the 3rd year of the study (Jensen et al., 2007). Current medication use did not account for this loss of relative benefit. Therefore, the accumulated evidence at this point in the series of reports of outcomes in the MTA suggests that the relative superiority of medication may be temporary for the majority even when treatment is continued, although it may continue to be increasingly effective for about one third of such children. It is possible that this loss may be due to how the treatment is managed in the community setting, which may be very different than management according to the MTA algorithm (i.e., with regular monthly visits to obtain information from parents, teachers, and the child and to make medication adjustments on the basis of this close monitoring of compliance and outcomes).

e. Does the selective treatment of the most difficult cases mask the beneficial effects of stimulant medication?

The hypothesis of selection bias was not supported when rigorously tested by applying propensity score analysis (Swanson, Hinshaw, et al., 2007). Although the preponderance of evidence so far from the MTA does not suggest that the loss of relative superiority was an artifact of comparison of biased subgroups, it remains plausible that some bias not yet evaluated may be present. We are still considering exploratory analyses of differences in outcome based on who is treated (determined by self-selection as individuals choose to maintain, stop, or start medication) and how they are treated (determined by community practitioners who may not continue the close monitoring and adjustment of medication). It is clear that additional studies are needed to characterize who starts and who stops treatment with medication, and for what reasons, during the course of long-term treatment.

f. Do some identifiable subgroups benefit from one treatment (either pharmacological or psychosocial) more than other subgroups?

The end-of-treatment moderator analyses generated hypotheses that anxiety and socioeconomic status may identify subgroups with enhanced response to behavior therapy, but this was not confirmed in the follow-up analyses. However, growth mixture model analyses of outcome through the 36-month assessment provided another way to form subgroups based on different outcome trajectories. One subgroup (34% of the participants) showed modest but monotonic improvement over time. In this subgroup, those consistently treated with medication compared to those not treated show significant initial benefit that increases over the 3 years of treatment (Swanson, Hinshaw, et al., 2007). However, analyses of MTA baseline measures did not reveal how to predict membership in this subgroup in advance, so the empirical approach of monitoring response over time is necessary to identify this subgroup.

2. Practical Suggestions and Future Directions

The initial findings of the MTA at the 14-month assessment (MTA Cooperative Group, 1999) clearly show long-term benefits of the stimulant medication for longer than 1 year, addressing a fundamental question that of the MTA. These findings were important, because at the time there was little information in the literature on the long-term effects of stimulants. Prior to the MTA, many published studies had documented short-term beneficial effects over a few hours, a few days, a few weeks, or a few months. The MTA articles added to the literature by extending the evaluation to the long-term—defined initially as a little longer than a year, when clear benefits were present.

The findings from the first follow-up (MTA Cooperative Group, 2004a, 2004b), which documented partial persistence of the relative superiority of the MTA medication algorithm, provided another contribution to the sparse literature on long-term effects of medication. The partial loss of effect must be interpreted in light of maintenance of treatment, which was reduced in the transition to treatment as usual in the community. Another long-term treatment study (Abikoff et al., 2004) showed continued full benefit for 2 years when treatment was monitored and adjustments were made to the medication regime over the entire time period.

The findings from the longer term follow-up (Jensen et al., 2007; Molina et al., 2007; Swanson, Elliott, et al., 2007; Swanson, Hinshaw, et al., 2007) provide additional guidance for evidence-based treatment that do not contradict the initial findings but do provide a new perspective. These additional findings suggest the possibility that the relative superiority of the intensive, carefully monitored medication management gradually dissipates when the children are returned to community treatment, and the longer term relative benefits are not apparent even when compared to children who are never treated with stimulant medication.

However, because the MTA outcomes are based on averages across individual children, it is likely that the relative superiority of medication will not dissipate in all cases, so individualization of long-term treatment guided by evaluation of withdrawal of medication makes sense.

Despite the challenges of integrating controversial findings into prior literature and practice, continued follow-up of the MTA participants could provide answers to a number of important questions. We have presented the outcomes at the 8-year follow-up (MTA Research Symposium, 2007), and the 10-year follow-up has been completed. The 12-year follow-up is now in progress. In the future, based on additional information from the MTA follow-up, we expect to address many new questions, including questions about access to services for children with ADHD, particularly for families seeking high-quality mental health treatment; about ultimate attainment of height and weight for children who show initial stimulant-related growth suppression; and about other long-term symptomatic and functional outcomes associated with ADHD.

In conclusion, the MTA continues to provide information that is both clinically relevant and intellectually stimulating. It is still providing evidence about long-term effects of treatment, generating interest in pertinent issues related to the long-term course of ADHD, and setting directions for future investigations. It will be crucial to examine further patterns of treatment-related effects during the current 12-year follow-up of participants in the MTA that is now in progress.

Acknowledgements

The Multimodal Treatment Study of Children with ADHD (MTA) was funded by the National Institute of Mental Health (NIMH) as a cooperative agreement involving six clinical sites. Collaborators from the National Institute of Mental Health are Peter S. Jensen, MD (currently at the REACH Institute), L. Eugene Arnold,MD, MEd

(currently at Ohio State University), Joanne B. Severe, MS (Clinical Trials Operations and Biostatistics Unit, Division of Services and Intervention Research), Benedetto Vitiello, MD (Child & Adolescent Treatment and Preventive Interventions Research Branch), Kimberly Hoagwood, PhD (currently at Columbia University); previous contributors from NIMH to the early phase: John Richters, PhD (currently at National Institute of Nursing Research); Donald Vereen, MD (currently at National Institute on Drug Abuse). Principal investigators and co-investigators from the clinical sites are from University of California, Berkeley/San Francisco: Stephen P. Hinshaw, PhD (Berkeley) and Glen R. Elliott, PhD, MD (San Francisco); Duke University: C. Keith Conners, PhD, Karen C. Wells, PhD, John March, MD, MPH, and Jeffery Epstein, PhD; University of California, Irvine/Los Angeles: James Swanson, PhD (Irvine), Dennis P. Cantwell, MD, (deceased, Los Angeles), and Timothy Wigal, PhD (Irvine); Long Island Jewish Medical Center/Montreal Children's Hospital: Howard B. Abikoff, PhD (currently at New York University School of Medicine) and Lily Hechtman, MD (McGill University); New York State Psychiatric Institute/Columbia University/Mount Sinai Medical Center: Laurence L. Greenhill, MD (Columbia University) and Jeffrey H. Newcorn, MD (Mount Sinai School of Medicine); University of Pittsburgh:William E. Pelham, PhD (currently at State University of New York, Buffalo), Betsy Hoza, PhD (currently at University of Vermont), and Brooke Molina, PhD The original statistical and trial design consultant was Helena C. Kraemer, PhD (Stanford University), and the follow-up phase statistical collaborators are Robert D. Gibbons, PhD (University of Illinois, Chicago), Sue Marcus, PhD (Mt. Sinai College of Medicine), and Kwan Hur, PhD (University of Illinois, Chicago). Additional collaborators are Thomas Hanley, EdD from the Office of Special Education Programs/US Department of Education and Karen Stern, PhD from the Office of Juvenile Justice and Delinquency Prevention/Department of Justice.

The work reported was supported by cooperative agreement grants and contracts from the National Institute of Mental Health to the following: University of California, Berkeley: U01 MH50461 and N01MH12009; Duke University: U01 MH50477 and N01MH12012; University of California, Irvine: U01 MH50440 and N01MH 12011; Research Foundation for Mental Hygiene (New York State Psychiatric Institute/Columbia University): U01 MH50467 and N01 MH12007; Long Island-Jewish Medical Center U01 MH50453; New York University: N01MH 12004; University of Pittsburgh: U01 MH50467 and N01 MH 12010; McGill University N01MH12008. The Office of Special Education Programs of the U.S. Department of Education, the Office of Juvenile Justice and Delinquency Prevention of the Justice Department, and the National Institute on Drug Abuse also participated in funding.

The opinions and assertions contained in this report are the private views of the authors and are not to be construed as official or as reflectingthe views of the Department of Health and Human Services, the National Institutes of Health, or the National Institute of Mental Health.

References

Abikoff, M., Hechtman, L., Klein, R. G., Weiss, G., Fleiss, K., Etcovitch, J., et al. (2004). Symptomatic improvement in children with ADHD treated with long-term methylphenidate and multimodal psychosocial treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 802-811.

Arnold, L. E., Abikoff, H. B., Cantwell, D. P., Conners, C. K., Elliott, G., Greenhill, L. L., et al. (1997a). National Institute of Mental Health collaborative multimodal treatment study of children with ADHD (MTA): Design challenges and choices. *Archives of General Psychiatry, 54*, 865-870.

Arnold, L. E., Abikoff, H. B., Cantwell, D. P., Conners, C. K., Elliott, G., Greenhill, L. L., et al. (1997b). National Institute of Mental Health collaborative multimodal treatment study of children with ADHD (MTA): Design, methodology, and protocol evolution. *Journal of Attention Disorders, 2*(3), 141-150.

Arnold, L. E., Chuang, S., Davies, M., Kraemer, H. C., Abikoff, H. B., Conners, C. K., et al. (2004). Nine months of multicomponent behavioral treatment for ADHD and effectiveness of MTA fading procedures. *Journal of Abnormal Child Psychology, 32*(1), 39-51.

Arnold, L. E., Elliott, M., Sachs, L., Bird, H., Kraemer, H. C., Wells, K. C., et al. (2003). Effects of ethnicity on treatment attendance, stimulant response/dose, and 14-month outcome in ADHD. *Journal of Consulting and Clinical Psychology, 71*, 713-727.

Barkley, R. A. (2000). Commentary on the multimodal treatment study of children with ADHD. *Journal of Abnormal Child Psychology*, *28*, 595-599.

Biederman, J., Wilens, T., Mick, E., Spencer, T. J., & Faraone, S. V. (1999). Pharmacotherapy of attention-deficit/hyperactivity reduces risk of substance use disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *29*, 546-557.

Boyle, M. H., Jadad, A. R., & Allnut, D. R. (1999). Lesson from large trials: The MTA study as a model for evaluating the treatment of childhood psychiatric disorder. *Canadian Journal of Psychiatry*, *44*, 991-998.

Breggin, P. R. (2001). MTA study has flaws. *Archives of General Psychiatry, 58*, 1184.

Conners, C. K., Epstein, J. N., March, J. S., Angold, A.,Wells, K. C., Klaric, J., et al. (2001). Multimodal treatment of ADHD in the MTA: An alternative outcome analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(2), 159-167.

Cunningham, C. E. (1999). In the wake of the MTA: Charting a new course for the study and treatment of children with attention-deficit hyperactivity disorder. *Canadian Journal of Psychiatry, 44*, 999-1006.

DeVeaugh-Geiss, J., March, J., Shapiro, M., Andreason, P. J., Emslie, G., Ford, L. M., et al. (2006). Child and adolescent psychopharmacology in the new millennium: A workshop for academia, industry, and government. *Journal of American Academic Child Adolescent Psychiatry*, *45*, 261-270.

Dopfner, M., Breuer, D., Schurmann, S., Metternich, T. W., Rademacher, C., & Lehmkuhl, G. (2004). Effectiveness of an adaptive multimodal treatment in children with attention-deficit/hyperactivity disorder—global outcome. *European Child Adolescent Psychiatry, 13*, 117-129.

Gibbons, R.D., Hedeker, D, Elkin, I, Waternaux, C, Kraemer, H.C., Greenhous, J.B., Shea, M.T., Imber, S.D., Sotsky, S.M., Watkins, J.T. Some conceptual and statistical issues in analysis of longitudinal psychiatric data: Applications to the NIMH treatment of depression collaboration research program dataset. *Archives of General Psychiatry, 50*, 739-750.

Greene, R. W., & Ablon, J. S. (2001). What does the MTA study tell us about the effective psychosocial treatment for ADHD? *Journal of Clinical Child Adolescent Psychology, 30*, 114-121.

Greenhill, L. L., Abikoff, H. B., Arnold, L. E., Cantwell, D. P., Conners, C. K., Elliott, G., et al. (1996). Medication treatment strategies in the MTA study: Relevance to clinicians and researchers. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 1304-1313.

Greenhill, L. L., Swanson, J. M., Vitiello, B., Davies, M., Clevenger, W., Wu, M., et al. (2001). Impairment and deportment responses to different methylphenidate doses in children with ADHD: The MTA titration. *Journal of the American Academy of Child and Adolescent Psychiatry 40*, 180-187.

Hechtman, L., Etcovitch, J., Platt, R., Arnold, L. E., Abikoff, H. B., Newcorn, J. H., et al. (2005). Does multimodal treatment of ADHD decrease other diagnoses? *Clinical Neuroscience Research, 5*(5-6), 273-282.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* New York: John Wiley & Sons.

Hinshaw, S. P. (2007). Moderators and mediators of treatment outcome for youth with ADHD: Understanding for whom and how interventions work. *Journal of Pediatric Psychology, Advanced Access*.

Hinshaw, S. P., March, J., Abikoff, H. B., Arnold, L. E., Cantwell, D. P., Conners, C. K., et al. (1997). Comprehensive assessment of childhood attention-deficit hyperactivity disorder in the context of a multisite, multimodal clinical trial. *Journal of Attention Deficit Disorders, 1*, 217-234.

Hinshaw, S. P., Owens, E. B., Wells, K. C., Kraemer, H. C., Abikoff, H. B., Arnold, L. E., et al. (2000). Family processes and treatment outcome in the MTA: Negative/ineffective parenting practices in relation to multimodal treatment. *Journal of Abnormal Child Psychology, 28*(6), 555-568.

Hoagwood, K., Jensen, P., Arnold, L. E., Roper, M., Severe, J., Odbert, C., et al. (2004). Reliability of the services for children and adolescents parent interview (SCAPI). *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(11), 1345-1354.

Jensen, P. S. (2001). ADHD comorbidity and treatment outcomes in the MTA. *Journal of the American Academy of Child and Adolescent Psychiatry, 40,* 134-136.

Jensen, P. S. (1999). Facts vs. fancies concerning the multimodal treatment study of attention deficit hyperactivity disorder (MTA). *Canadian Journal of Psychiatry*, *44*, 975-980.

Jensen, P. S., Arnold, L. E., Swanson, J., Vitiello, B., Abikoff, H. B., Greenhill, L. L., et al. (2007). Follow-up of the NIMH MTA study at 36 months after randomization. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*(8), 988-1001.

Jensen, P. S., Hinshaw, S. P., Kraemer, H. C., Lenora, N., Newcorn, J. H., Abikoff, H. B., et al. (2001). ADHD comorbidity findings from the MTA study: Comparing comorbid subgroups. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(2), 147-158.

Jensen, P. S., Hinshaw, S. P., Swanson, J. M., Greenhill, L. L., Conners, C. K., Arnold, L. E., et al. (2001). Findings from the NIMH Multimodal Treatment Study of ADHD (MTA): Implications and applications for primary care providers. *Developmental Behavior Pediatrics, 22*, 60-73.

Jensen, P. S., Hoagwood, K., & Trickett, E. J. (1999). Ivory towers or earthen trenches? Community collaborations to foster real-world research. *Applied Developmental Science*, *3*, 206-212.

Klein, R. G. (2001). MTA findings fail to consider methodological flaws. *Archives of General Psychiatry, 58*, 1184-1185.

Kraemer H.C., Yesavage, J.A., Taylor, J.L., & Kupfer, D. (2000). How can we learn about developmental processes from cross-sectional studies, or can we? *American Journal of Psychiatry*, *157*(2): 163-171.

Kraemer H.C., & Robinson, T.N. (2005). Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? *Contemporary Clinical Trials*, 26: 518-529.

Kraemer, H. C.,Wilson, T. G., Fairburn, C. G., & Agras,W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59*, 877-883.

Marcus, S. M., & Gibbons, R. D. (2001). Estimating the efficacy of receiving treatment in randomized clinical trials with noncompliance. *Health Services and Outcomes Research Methodology, 2*, 247-258.

Molina, B. S. G., Flory, K., Hinshaw, S. P., Greiner, A. R., Arnold, E., Swanson, J., et al. (2007). Delinquent behavior and emerging substance use in the MTA at 36-months: Prevalence, course, and treatment effects. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*(8), 1027-1039.

MTA Cooperative Group. (2004a). National Institute of Mental Health multimodal treatment study of ADHD follow-up: Changes in effectiveness and growth after the end of treatment. *Pediatrics, 113*(4), 762-769.

MTA Cooperative Group. (1999a). 14-month randomized clinical trial of treatment strategies for attention deficit hyperactivity disorder. *Archives of General Psychiatry, 56*, 1073-1086.

MTA Cooperative Group. (1999b). Moderators and mediators of treatment response for children with attention-deficit/hyperactivity disorder: The multimodal treatment study of children with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry, 56*, 1088-1096.

MTA Cooperative Group. (2004b). National Institute of Mental Health multimodal treatment study of ADHD follow-up: 24-month outcomes of treatment strategies for attention-deficit/hyperactivity disorder (ADHD). *Pediatrics, 113*(4), 754-761.

MTA Research Symposium. (2007). *Investigators' opinions of important findings*. Washington DC: CHADD.

National Institute of Health. (1998, November 16-18). *Diagnosis and treatment of attention deficit hyperactivity disorder*. Retrieved from http://consensus.nih.gov/1998/1998AttentionDeficitHyperactivityDisorder110htmlhtm

National Institute of Health. (2005, November 14-15). *Program on clinical research policy analysis and coordination considering usual medical care in clinical trial design: Scientific and ethical issues*. Bethesda, MD: Author.

Owens, E. B., Hinshaw, S. P., Kraemer, H. C., Arnold, L. E., Abikoff, H. B., Cantwell, D. P., et al. (2003). Which treatment for whom for ADHD? Moderators of treatment response in the MTA. *Journal of Consulting and Clinical Psychology, 71*(3), 540-552.

Pelham, W. E. (1999). The NIMH multimodal treatment study for ADHD: Just say yes to drugs alone? *Canadian Journal of Psychiatry, 44*(10), 981-990.

Pelham,W. E., Gnagy, E. M., Greiner, A. R., Hoza, B., Hinshaw, S. P., Swanson, J. M., et al. (2000). Behavioral vs. behavioral and pharmacological treatment in ADHD children attending a summer treatment program. *Journal of Abnormal Child Psychology, 28*(6), 507-526.

Richters, J. E., Arnold, L. E., Jensen, P. S., Abikoff, H., Conners, C. K., Greenhill, L. L., et al. (1995). NIMH collaborative multisite multimodal treatment study of children with ADHD: I: Background and rationale. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*, 987-1000.

Roche, A. F., Lipman, R. S., Overall, J. E., & Hung, W. (1979). The effects of stimulants medication of the growth of hyperkinetic children. *Pediatrics*, *63*, 847-850.

Safer, D. J., Allen, R. P., & Barr, E. (1975). Growth rebound after termination of stimulant drugs. *Journal of Pediatrics*, *86*, 113-116.

Santosh, P. J., Taylor, E., Swanson, J.,Wigal, T., Chuang, S., Davies, M., et al. (2005). Reanalysis of the multimodal treatment study of attentiondeficit/hyperactivity disorder (ADHD) based on ICD-10 criteria for hyperkinetic disorder (HD). *Clinical Neuroscience Research, 5*(5-6).

Spencer T, Biederman J, Harding M, O'Donnell D, Faraone SV, Wilens TE (1996), Growth deficits in ADHD children revisited: evidence for disorder-associated growth delays? *J. Am. Acad. Child Adolesc. Psychiatry 35*, 1460-1469.

Swanson, J. M. (2005, November 14-15). *Case presentation: Case study #2: Multimodal treatment study of ADHD (MTA): Considering usual medical care in clinical trial design*. NIH Program on Research Policy Analysis and Coordination, Bethesda, MD.

Swanson, J. M., Arnold, L. E., Vitiello, B., Abikoff, H. B.,Wells, K. C., Pelham,W. E., et al. (2002). Response to commentary on the multimodal treatment of study of ADHD (MTA): Mining the meaning of the MTA. *Journal of Abnormal Child Psychology, 30*, 327-332.

Swanson, J. M., Elliott, G. R., Greenhill, L. L.,Wigal, T., Arnold, L. E., Vitiello, B., et al. (2007). Effects of stimulant medication on growth rates across 3 years in the MTA follow-up. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*(8), 1014-1026.

Swanson, J. M., Hinshaw, S. P., Arnold, L. E., Gibbons, R., Marcus, S., Hur, K., et al. (2007). Secondary evaluations of MTA 36-month outcomes: Propensity score and growth mixture model analyses. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*(8), 1002-1013.

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P.,Arnold, L. E., Conners, C. K., Abikoff, H. B., et al. (2001). Clinical relevance of the primary findings of the MTA: Success rates based on severity of ADHD and ODD symptoms at the end of treatment. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(2), 168-179.

Taylor, E. (1999). Development of clinical services for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1097-1099.

van der Oord, S., Prins, P. J., Oosterlaan, J., & Emmerkamp, P. M. (2007). Does brief, clinically based, intensive multimodal behavioral therapy enhance the effects of methylphenidate in children with ADHD? *European Child Adolescent Psychiatry*, *16*, 48-57.

Vitiello, B., Severe, J. B., Greenhill, L. L., Arnold, L. E., Abikoff, H. B., Bukstein, O., et al. (2001). Methylphenidate dosage for children with ADHD over time under controlled conditions: Lessons from the MTA. *Journal of the American Academy of Psychiatry, 40*, 188-196.

Wells, K. C. (2001). Comprehensive versus matched psychosocial treatment in the MTA study: Conceptual and empirical issues. *Journal of Clinical Child Psychology, 30*, 131-135.

Wells, K. C., Chi, T. C., Hinshaw, S. P., Epstein, J. N., Pfiffner, L. J., Nebel-Schwain, M., et al. (2006). Treatment-related changes in objectively measured parenting behaviors in the multimodal treatment study of children with ADHD. *Journal of Consulting and Clinical Psychology, 74*, 649-657.

Wells, K. C., Epstein, J., Hinshaw, S., Conners, C. K., Abikoff, H. B., Abramowitz, A., et al. (2000). Parenting and family stress treatment outcomes in attention deficit hyperactivity disorder(ADHD): An empirical analysis in the MTA study. *Journal of Abnormal Child Psychology, 28*(6), 543-554.

Wells, K. C., Pelham, W. E., Kotkin, R. A., Hoza, B., Abikoff, H. B., Abramowitz, A., et al. (2000). Psychosocial treatment strategies in the MTA study: Rationale, methods, and critical issues in design and implementation. *Journal of Abnormal Child Psychology, 28*(6), 483-506.

Wilens TE, Faraone SV, Biederman J, Gunawardene S: Does stimulant therapy of attention-deficit/hyperactivity disorder beget later substance abuse? A meta-analytic review of the literature. *Pediatrics* 2003; 111:179-185.