THE FINDINGS AND CONCLUSIONS IN THIS REPORT HAVE NOT BEEN FORMALLY DISSEMINATED BY FDA AND SHOULD NOT BE CONSTRUED TO REPRESENT ANY AGENCY DETERMINATION OR POLICY.

Addendum Review to the ORNL IAG Project Assignment (#2008-31) on "*The Evaluation of proposed Association Between Artificial Food Colors and Attention Deficit Hyperactivity Disorders (ADHD) and Problem Behaviors in Children*"

Thomas J. Sobotka, Ph.D.
August 25, 2010

Table of Contents

I INTRODUCTION 2
II REVIEW and EVALUATION OF *SCHAB and TRINH (2004)* 2
A. Study Rationale and Design 2
B. Study Results and Conclusions 3
C. Assessment and Summary Evaluation of Study 5

I INTRODUCTION

The present review and evaluation report serves as an addendum to the ORNL IAG project assignment (#2008-31) on "*The Evaluation of proposed Association Between Artificial Food Colors and Attention Deficit Hyperactivity Disorders (ADHD) and Problem Behaviors in Children*". The subject of this review/evaluation is a study conducted by *Schab and Trinh (2004*) which presents the results of a meta-analysis of findings from previously conducted clinical trials addressing the proposed relationship between consumption of artificial food colors (AFCs) and behavioral changes in children diagnosed with hyperactivity (ADHD). This meta-analysis has been cited by the Center for Science in the Public Interest (CSPI; Petition submitted to the FDA on June 3, 2008; File FDA.2008.!0349.0001) as a pivotal study in CSPI's assertion that certified color additives cause adverse behavioral effects in some children; therefore, it was deemed important that a thorough evaluation of this study should be included in FDA's review of the literature.

II REVIEW and EVALUATION OF *SCHAB and TRINH (2004)*

Schab DW Trinh NT. Do artificial food colors promote hyperactivity in children with hyperactive syndromes? A meta-analysis of double-blind placebo-controlled trials. J Dev Behav Pediatr. 2004 (December), 25(6):423-34.

A. Study Rationale and Design

The study conducted by Schab and Trinh (2004) attempted to provide a focused evaluation of whether artificial food colors (AFC) promote the symptomatology of hyperactivity in children diagnosed with hyperactive syndromes, as measured on behavioral rating scales. These authors conducted a primary meta-analysis of findings, based only on behavior ratings, from 15 selected previously conducted double-blind placebo-controlled clinical trials (total of 219 subjects) addressing a proposed relationship between consumption of AFCs and behavioral changes in children (7.9 years average age; male-female ratio at 5.5 to 1) with a diagnosis of hyperactivity. It should be emphasized that, other than the behavioral ratings (parent, teacher, clinician), this meta-analysis did not consider any of the more objective outcome measures of behavior such as clinical/psychological evaluations, activity monitoring, and behavioral testing which were available in many of the trials. The same study data as used in the primary meta-analysis were also analyzed to address three sub-hypotheses that may help explain the heterogeneity of prior trials' results, facilitate identification of responders to AFCs, or aid in design of future trials: (1) *Sub-hypothesis 1*: parents and teachers differ in their reports of responsiveness to AFCs; (2) *Sub-hypothesis 2:* open trial and parental report can be used as screening methods to identify potentially responsive children; and (3) *Sub-hypothesis 3:* children formally diagnosed with hyperactivity have greater responsiveness to AFCs than do heterogeneous or non-hyperactive groups of children. Additionally, a secondary analysis was conducted to address the three sub-hypotheses using data from an additional 8 clinical trials (total of 132 subjects) which met the primary inclusion criteria except that the subject population consisted of either exclusively non-hyperactive or a heterogeneous mixture of hyperactive and non-hyperactive children (7.3 years average age; male-female ratio at 3.5 to 1). To standardize the behavior rating data across clinical trials, the authors used a random effects model with the standardized mean difference

(SMD) as the measure of effect size (ES). The SMD describes the difference in outcome between the active and control arms of a trial in terms of the number of pooled standard deviations by which the two groups differ. The summary ES depicts the summary response to AFCs. Adverse effects on hyperactivity are reflected in positive ES values.

B. Study Results and Conclusions

Primary Analysis w Hyperactive Children:

Based on the behavioral ratings of parents, teachers and clinicians combined from 15 clinical trials meeting the primary inclusion criteria with children diagnosed with hyperactivity syndrome, meta-analytic modeling identified a statistically significant overall summary response to artificial food colors (AFCs) with an effects size (ES) of 0.283 (95% CI, 0.079 to 0.488), reflecting a change of slightly more than 1/4 of a SD. The authors state that the magnitude of this ES is approximately 1/3 to 1/2 the magnitude of the ES calculated in a meta-analysis of trials evaluating methylphenidate as a treatment of ADHD (Schachter et al, 2001). The authors conclude that these results suggest an association between ingestion of AFCs and problem behaviors in hyperactive children. The need for qualifications to this conclusion is explained below in the section "Assessment and Evaluation of the Study".

Testing the *sub-hypothesis 1* showed that, while the ES values for teachers (ES = 0.0810; 95% CI, -0.073 to 0.235) and clinicians (ES = 0.107; 95% CI, -0.128 to 0.343) were not statistically significant, the ES values for parents (ES = 0.441; 95% CI, 0.161 to 0.721) were statistically significant. However, since no statistical difference was found between these three groups' scores and no relevant information is provided in the secondary analysis, the authors concluded that the *first sub-hypothesis*, that parents, teachers and clinicians differ in their reports of hyperactivity behaviors, could not be confirmed. Although this latter conclusion may be statistically correct, it is practically incongruous with the fact that, as the authors state, "only the parents' summary ES was statistically significant". Effectively, as the authors also state, "in our results, parents tended to detect greater change from the intervention [AFCs] than did teachers [or clinicians]". As discussed below in the section "Assessment and Evaluation of the Study", this issue of parents detecting behavioral changes not noticed by other raters is relevant to interpretation of the results from this meta-analysis.

Testing the effectiveness of screening hyperactive children as potentially responsive to AFCs (*sub-hypothesis 2*) revealed that trials that did not prescreen for potential responders demonstrated a nonsignificant effect of AFCs on behavior ratings in controlled testing (ES = 0.09; 95% CI, -0.108 to 0.288) but trials that selected subjects who were prescreened as presumptively responsive to AFCs did demonstrate a statistically significant effect of AFCs (ES = 0.535; 95% CI, 0.149 to 0.920). However, there was no statistically significant difference between these two groups. The significance of these findings will be discussed in conjunction with the results from the secondary analysis below.

As a partial approach toward addressing the *third sub-hypothesis* that hyperactive children have greater responsiveness to AFCs than non-hyperactive children or children

heterogeneous for hyperactivity, the trials in the primary analysis which used only hyperactive children were ordered according to diagnostic grades, i.e. severity of hyperactivity. No differential trends in ES values were found across these groupings of trials. The significance of these findings will be discussed in conjunction with the results from the secondary analysis below.

Secondary Analysis w Non-Hyperactive or Heterogeneous Groups of Children: Based on 8 clinical trials (secondary trials) meeting the primary inclusion criteria with non-hyperactive children or groups of children heterogeneous for hyperactivity, meta-analysis revealed a non-significant ES of 0.117 (95% CI, -0.113 to 0.347) indicating no summary effect of AFCs on behavior ratings for this group of trials.

In the absence of a significant ES, the secondary trials were not tested for *sub-hypothesis 1* (i.e., difference between parents' and teachers' ratings).

However, when the secondary trials were analyzed separately for those that pre-screened or did not pre-screen for potential responsiveness to colors (*sub-hypothesis 2*), there was a significant difference between these two groups. Trials that did not pre-screen showed a non-significant response to AFCs in controlled testing (ES = -0.112; 95% CI, -0.393 to 0.169); however, a statistically significant ES of 0.316 (95% CI, 0.157 to 0.475) characterized the trials that did pre-screen. The difference in ESs between these secondary trials that prescreened and did not prescreen children was statistically significant. Post-hoc analysis combining the primary and secondary trials also showed that, among all trials, those that prescreened were significantly different from those that did not screen. Based on both the primary and secondary analyses, the authors concluded that the *second hypothesis*, that children potentially responsive to AFCs can be identified through screening methods, was only partially substantiated. Overall, trials that pre-screened for responsiveness demonstrated a significant treatment effect in controlled testing whether enrolling hyperactive (primary trials) or heterogeneous, predominantly non-hyperactive subjects. However, there was a statistically significant difference between trials that pre-screened and those that did not only among the secondary group of trials enrolling heterogeneous, predominantly non-hyperactive subjects. Interestingly, the authors suggested that "screening may be more effective in heterogeneous populations because of greater variability in responsiveness to AFCs among such populations than among the strictly hyperactive populations". This reviewer suggests the possibility that this unique responsiveness to AFCs may be associated with a genetic polymorphism that may be more variable in the heterogeneous populations than in the strictly hyperactive populations (refer to Color/ADHD Overview Report by Sobotka, 2010).

In testing the *third sub-hypothesis*, that hyperactive children have greater responsiveness to AFCs than non-hyperactive children or children heterogeneous for hyperactivity, the summary ES of the trials in the secondary analysis, whose participants were predominantly non-hyperactive (but with other problem behaviors), did not differ significantly from the summary ES of the trials in the primary analysis (hyperactive children). The authors concluded that this negates the *third sub-hypothesis* and shows that presence of a baseline diagnosis of hyperactivity and rigor of that diagnosis does not correspond to reactivity to AFCs, suggesting that AFC-sensitivity among patients may not be limited to those with clear-cut criteria-specific hyperactive syndrome.

C. Assessment and Summary Evaluation of Study

One prefacing comment relevant to assessing this study and its conclusions is that one important element in the conduct of a meta-analysis involves consideration of the quality of the various trials included in the analysis and the confidence in the data and findings from those trials. According to the authors of this study, the various trials used in this meta-analysis were "graded according to the degree that they adequately addressed types of bias and avoided important threats to validity: differential selection, performance, attrition, detection, and reporting". In the present study (page 426, section "*Primary Analysis*") the various trials were assigned "validity" scores of "A" (the highest score) to "C" (the lowest). Although the scoring criteria were not specified and the individual scores for all trials were not presented, among the 15 trials used for the Primary Analysis, 2 trials received the highest "A" validity score, 2 trials received the lowest "C" score, and apparently most of the trials (11) were scored as having a moderate "B" level of validity. It is unclear, however, how or whether the weighting values of these "validity" scores and other experimental limitations were actually factored into the basic primary meta-analysis in this study. The only described use of the "validity" scores, as indicated on page 428 of the document (section "*Sensitivity Analysis of Primary Trials"*), was in the conduct of an additional separate meta-analyses which excluded 2 trials with "C" validity scores. This separate analysis reduced the ES (Effect Size) from the summary primary meta-analysis of 0.283 (statistically significant) to a recalculated ES of 0.216 (statistically significant). An additional separate meta-analysis was conducted to exclude 2 trials with the smallest numbers of subjects from the analysis, reducing the ES score to 0.255 (statistically significant). Analysis with both types of trials excluded (lowest quality and smallest) yielded a summary ES score of 0.210 (still statistically significant). Since there is no other mention of the "validity" scores being used in this analysis, it appears, therefore, that the primary meta-analysis (ES = 0.283) in this study was conducted with findings from the various trials that were used with little adjustment of the analysis to factor in measure(s) of quality of the clinical trials and confidence in the reported findings from those trials. However, in the Discussion section of this paper the authors do attempt to explain the significance of many of the experimental design issues and other shortcomings of the trials used for this meta-analysis and their impact on interpretation of this study's results. The authors also include recommendations whereby future research may avoid the various problematic experimental methods and design issues which many prior trials have experienced. This reviewer suggests that input from an experienced statistician would be beneficial in providing a more cogent assessment of the appropriateness of the statistical procedures used in the conduct of this meta-analysis, including the significance of the "validity" scores in accounting for the quality of the trials and confidence in their findings.

The authors open their discussion of their study's findings with the statement that this "meta-analysis supports the hypothesis that AFCs promote hyperactivity in hyperactive children, as measured on behavioral rating scales". This reviewer considers this a rather broad overstatement of the significance of their findings. Later in the discussion the authors do state appropriately that "several limitations, especially of the included trials, but also of our meta-analytic methods, restrict the interpretation of our results" and conclude more correctly that their results only "strongly suggest an association of AFCs

and hyperactivity". Among the limitations referred to by the authors were the use of unorthodox scales to measure behavioral responses which may limit the validity of conclusions about the effects of AFCs on hyperactivity (but suggest that AFCs provoke general behavioral disturbance more than hyperactivity), inappropriate statistical methods in many of the trials, clinical heterogeneity of the study population together with the trials' inconsistent reporting of subtypes, and the possibility of that this study's use of only behavioral rating outcomes may make this study irrelevant to ADHD's inattentive subtype, whose symptoms may not be predominantly behavioral phenomena. The latter limitation, that is the use of only behavioral rating outcomes for this meta-analysis, has a broader relevance to the interpretation of the results of this study particularly in view of the results of testing the *second sub-hypothesis* in this study that compared the sources of the behavioral ratings, i.e. parents, teachers, and clinicians. The outcome of testing the *second sub-hypothesis* showed that only the summary ES of parents' ratings was statistically significant but neither the teacher based nor clinician based ESs were significant (although there was no statistical difference between the three groups' scores). This appears to indicate that the parental behavioral ratings may be the major contributor to the overall significant ES associated with AFCs and that the teacher and clinician ratings provide little if any support for a behavioral response to AFCs. Several trials have reported similar findings with parent ratings showing effects associated with AFCs but teacher or clinician ratings providing contradictory or otherwise no supporting confirmation (Harley et al, 1979; Mattes and Gittelman-Klein, 1981; Williams et al, 1978). In discussing the *second sub-hypothesis* finding that parents may have detected behavioral change unnoticed by teachers and clinicians, the authors considered several possible reasons for this difference, including the following:

1. *that parents were ineffectively blinded* – The authors cited 5 trials (Goyette et al, 1978a and b; Harley et al, 1978; Mattes and Gittelman-Klein, 1978; Rose, 1977) as having conducted rigorous testing and confirmed the effectiveness of the blind for parents, subject and investigators (2 of the cited trials [Goyette et al, 1978 a and b] in fact provided no description of the blinding procedures used in their trials). This reviewer considers this less than convincing evidence to conclude that the parents were effectively blinded in the majority of trials used for the meta-analysis and to dismiss the possibility that ineffective blinding may have unintentionally resulted in the parents reporting behavioral changes that were not evident to either the teachers or clinicians. The authors even include "explicitly testing the blinding of subjects" as one of the recommendations whereby future research may avoid the various problematic experimental methods and design issues which many prior trials have experienced.
2. *that parents' based their ratings on problem behaviors of concern to them (irritability and sleep disturbance) which may differ from those of concern to teachers and clinicians (restlessness and inattention) and that AFCs may promote a pattern of symptoms that is incongruent with modern criteria of ADHD but that is nevertheless bothersome to parents* – this reflects suggestions made by several investigators the effects of AFCs may not be associated with ADHD types of behavior but more with general problem behaviors (Carter et al, 1993; Mattes and Gittelman-Klein, 1978; Rowe, 1988; Rowe and Rowe, 1994; Sarantinos et al, 1990).
3. *parents may be particularly attuned to the idiosyncrasies of their own children* – several investigators attempted to address this possibility by using specially developed questionnaires specifically compiled by parents (Weiss et al, 1980) or

investigators (Rowe, 1988) to target the unique behaviors of the subject children. Neither study identified a higher percent of responders than trials using standard rating questionnaires indicating that this is not a likely reason for parents to detect behavioral changes not evident to other raters.

In the absence of a suitable explanation for the detection differences between parent ratings and teacher and clinician ratings, this reviewer considers that parental reports of behavioral changes unconfirmed or not support by other sources or measures of behavioral change should generally be viewed as inconclusive and only suggestive of a treatment effect.

Overall, the primary results of this meta-analysis of behavioral rating findings from 15 controlled clinical trials with diagnosed hyperactive children did show a small but significant summary effect of AFCs on parental ratings, but no significant treatment effect based on either teacher or clinician ratings. In view of the absence of confirmation of treatment effects with either the teacher or clinician behavioral ratings and the questionable use of the “validity” scores for the trials used in this analysis, the significance of the effects identified in this meta-analysis is viewed as inconclusive but suggestive of an association between AFCs and hyperactivity behaviors. This interpretation may be considered consistent with the fact that the findings from the various clinical trials used in this analysis were generally considered only suggestive of a limited association between AFCs and adverse behaviors in children with hyperactivity or other problem behaviors. Importantly, the findings from this meta-analysis do not negate either the general view that only a small subgroup of children appears to be affected by AFCs or the suggestion that these responsive children may be among a subgroup of children who are uniquely sensitive (intolerant) to a variety of food related chemicals not limited to AFCs which may adversely affect behavior in this select subgroup of children. Additionally, the findings from this study do highlight questions other investigators have raised about the specificity or nature of effects associated with AFCs, specifically that the symptomatology of ADHD may differ from the pattern of symptoms induced by AFCs, the latter being associated more with provoking general behavioral disturbances such as irritability and insomnia than with eliciting ADHD symptoms such as restlessness and inattention (Carter et al, 1993; Mattes and Gittelman-Klein, 1978; Rowe, 1988; Rowe and Rowe, 1994; Sarantinos et al, 1990). The present authors suggest a broader focus such that "researchers should not limit their search for responders to those previously diagnosed with hyperactivity" and thereby may progress "out of the realm of hyperactivity and into a wider province of behavioral pathology".