
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.

STUDY EVALUATION REPORT

Chronic and Acute Effects of Artificial Colourings and Preservatives on Children's Behaviour

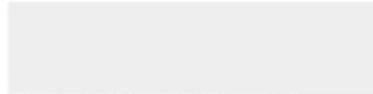
STUDY TYPE: Clinical Study

University of Southampton School of Psychology
Project Code: T07040

Prepared for
Center for Food Safety and Applied Nutrition
Office of Food Additive Safety
U.S. Food and Drug Administration
College Park, MD 20740-3835
Task Order (Assignment) No. 2008-32

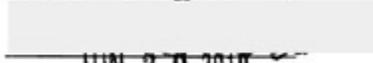
Prepared by
Toxicology and Hazard Assessment Group
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831

Primary Reviewer:
Tom Sobotka, Ph.D.

Signature: 

Date: 6-29-10

Secondary Reviewers:
Carol Wood, Ph.D., D.A.B.T.

Signature: 

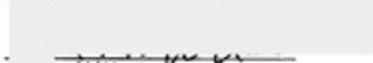
Date: JUN 29 2010

Robert H. Ross, M.S., Group Leader

Signature: 

Date: JUN 29 2010

Quality Assurance:
Lee Ann Wilson, M.A.

Signature: 

Date: JUN 29 2010

TABLE OF CONTENTS

LIST OF TABLES	4
Evaluation Report	5
I. Executive Summary	5
II. Introduction – Study Objectives	7
III. Study Design/Methods	7
A. Primary Study	7
1. Participants.....	7
2. Primary Study Design.....	7
3. Challenge Treatments.....	8
4. Challenge Protocol.....	9
5. Behavioral Measures.....	9
6. Data Analysis Methods.....	10
B. Additional Experimental Design Procedures to Address Salient Correlative Secondary Research Questions	11
1. Effects of the AFCA challenge mixes on component behavior measures of GHA.....	11
2. Behavioral effects and metabolic mediators following acute challenge with AFCA.....	11
3. Modulation of AFCA behavioral effects by genetic polymorphisms.....	12
C. Reviewer Comments Regarding Study Design/Methods	13
1. <i>Questionable accuracy of teacher pre-study behavioral ratings recalling behavior over a period of 6 months</i>	13
2. <i>Questionable use and accuracy of pre-study behavioral ratings by participating parents</i>	13
3. <i>Disproportionate increases were made in levels of artificial colors in the mixes (A and B) used for the 8/9YO children relative to the mixes (A and B) used for the 3/4YO children, and no adjustments were made to levels of sodium benzoate in either Mix A or B for either age group of children</i>	13
4. <i>Composition of placebo is unclear</i>	13
5. <i>Parents were not asked about the blinding</i>	14
6. <i>Timing of juice consumption at home</i>	14
7. <i>Precision of recall data in conduct of weekly parent/teacher ratings and availability of standard ratings</i>	14
8. <i>Reliability of modified Classroom Observation Code (COC) used in 3/4YO children is unclear</i>	14
9. <i>Questionable adequacy of duration used for classroom observation</i>	14
10. <i>Model 1 is not useable for analysis of data</i>	14
11. <i>In the “acute challenge phase” of this study the disaggregated behavior measures comprising the Hyperactivity Index (HI) were not analyzed separately</i>	15
12. <i>The Mixed Model analyses of the “acute challenge” data did not appear to control for potential confounding factors</i>	15
13. <i>No description of materials or methods was provided for the genotyping analyses</i>	15
IV. Study Investigators’ Reported Findings	15
A. Reported Study Findings Relative To Primary Research Question: Do Mixtures of Certain Artificial Food Colors and Sodium Benzoate Increase the Mean Level of Hyperactive Behavior (Hyperactivity Profile Behaviors) In Children From the General Population?	15
1. Study Sample and Background Characteristics.....	15
2. Juice consumption and dietary infractions.....	16
3. Effects of Challenge.....	16
B. Reviewer Comments Regarding Reported Study Findings for Primary Research Question	18
1. <i>Limited information regarding CPT performance decrements of 8/9YO children over study period</i>	18
2. <i>Problematic aspects of the data analyses</i>	18

3. <i>There is no clarification of the behavioral changes in this study as necessarily “adverse” (p.22) or “deleterious” (p.26)</i>	19
C. Reported Study Findings Regarding Secondary Research Questions	19
1. (Secondary Research Question) Effects of the AFCA challenge mixes on component behavior measures of GHA	19
2. <i>Reviewer Comments: Effects of the AFCA challenge mixes on component behavior measures of GHA</i>	21
(a) <i>Complementary changes across disaggregated behavior measures can provide reliable confirmation of significant treatment effects.</i>	21
(b) <i>Use of the three sample groups of children for analysis of the component (disaggregated) behavior measures results in variable statistical outcomes but relevance of this to data interpretation are not discussed</i>	21
3. (Secondary Research Question) Behavioral effects and metabolic mediators following acute challenge with AFCA – Study Findings and Interpretation	22
4. <i>Reviewer Comments: Behavioral effects and metabolic mediators following acute challenge with AFCA – Study Findings</i>	22
(a) <i>Statistical analyses of data in the “acute challenge” component study did not control for confounding variables.</i>	22
(b) <i>Rationale for explaining why treatment effects in the main study were detected only in parental ratings was based on inaccurate interpretation of acute study findings.</i>	23
5. (Secondary Research Question) Modulation of AFCA behavioral effects by genetic polymorphisms – Study Findings	23
6. <i>Reviewer Comments: Modulation of AFCA behavioral effects by genetic polymorphisms – Study Findings</i>	26
(a) <i>Variability in finding of significance for main effect of challenge (Mix A and Mix B) on behavioral responses (GHA) with repeated Mixed Models analyses for different genotypes raises questions about replicability of the Mixed Models analyses and uncertainties about the findings of significant treatment effects in the primary study.</i>	26
(b) <i>The moderating influence specifically of the Thr105Ile/present genotype decreasing the behavioral effects of Mix A in 3/4 YO children appears contradictory to the main effect of genotype on GHA levels.</i>	27
(c) <i>The findings from the analysis of the genotype component study data which involved children from the general population were inappropriately extrapolated to ADHD children.</i>	27
7. Additional Secondary Questions – Study Findings	27
(a) <i>Consistency in response between Mix A and Mix B.</i>	27
(b) <i>Dose-response relationship.</i>	28
(c) <i>Difference between the 3/4 year old and 8/9 year old children in changes in GHA over study period.</i> 28	
V. Study Investigators’ Overall Conclusions of Study	28
VI. Reviewer Conclusions	29
VII. REVIEWER NOTED STUDY STRENGTHS	32
VIII. REVIEWER NOTED STUDY WEAKNESSES	32
A. Procedural Weaknesses	33
B. Weaknesses Regarding Data Analyses and Interpretation of Study Findings	34
IX. Applicability to Assess Risk or to Support Regulatory Action	37
References	38

LIST OF TABLES

TABLE 1. Composition of Challenge Material 8
TABLE 2. Numbers of 3/4-Year Old and 8/9-Year Old Children in Study 15
TABLE 3. Unadjusted Mean GHA Scores* for 3/4YO and 8/9YO Children by Challenge Type
..... 16
TABLE 4. General Linear Mixed Models estimates of GHA effect size for challenge Mix A
versus Placebo and challenge Mix B versus Placebo 17
TABLE 5. Effect of sizes of Mix A versus Placebo and Mix B versus Placebo for each
disaggregated behavior measure 20
TABLE 6. Mixed Models analyses of interactions between Genotype and Behavioral Responses
(GHA) to Additive Mix A and Mix B versus Placebo in the $\geq 85\%$ consumption sub-
sample of 3/4 YO Children with All Potential Confounds Controlled 24
TABLE 7. Mixed Models analyses of interactions between Genotype and Behavioral Responses
(GHA) to Additive Mix A and Mix B versus Placebo in the $\geq 85\%$ consumption sub-
sample of 8/9 YO Children with All Potential Confounds Controlled 25

Evaluation Report

Project Title: Chronic and acute effects of artificial colourings and preservatives on children's behaviour

Principle Investigators: Jim Stevenson (PI)
Edmund Sonuga-Barke (co-PI)
John Warner (co-PI)

Contractor: School of Psychology, University of Southampton (England)

Study Dates: September 1, 2004 to February 28, 2007

Study Technical Report Date: Submitted June 18, 2007; Revised September 10, 2007

The format for this evaluation of the Southampton study on artificial colors/preservatives and children's behavior includes an Executive Summary followed by several sections summarizing information provided by the study investigators regarding: the study objectives; the study design and methods; the reported study findings (addressing the primary and secondary research questions); the investigators' overall conclusions; and the reviewer conclusions on the study. Each of these sections will include "reviewer" comments, specifically addressing the information in that section. The final three sections of this report will address: the strengths of the study, study weaknesses, and the applicability of the study findings for assessing risk or supporting regulatory action.

I. Executive Summary

The principle investigators pointed out that there is a longstanding suggestion, initiated by Ben Feingold (1975) more than 30 years ago, that artificial food colors and additives (AFCA), including preservatives, have detrimental effects on children, inducing an adverse level of overactive, impulsive and inattentive behaviors, i.e. "hyperactivity". Children who show this behavior pattern to a marked degree are also likely to be diagnosed with attention deficit hyperactivity disorder (ADHD). Although similar types of behaviors also occur among the general population of children, it is important to emphasize the distinction between the mild nature of these behaviors in the normal population (which is the focus of the present study) versus the severe, persistent and disruptive nature of these types of behaviors which characterize the abnormalities of ADHD. Earlier studies have failed to confirm the suggested causative association between AFCA and ADHD in children, although a recent meta-analysis of double-blind, placebo-controlled clinical studies reported a significant effect of AFCA on the behavior of children with ADHD. Whether AFCAs have a similar effect in the general population has not been conclusively demonstrated. The principle investigators of the present study conducted a previous study which provided some suggestive evidence of effects on *hyperactivity profile behaviors* based on parental ratings for 3 year old children from the general population in response to a mix of AFCA, but these findings were not replicated by concomitant clinical behavioral assessments.

The primary hypothesis tested in the present study was that mixtures of certain artificial food

colors and sodium benzoate (AFCA) increase the mean level of *hyperactivity profile behaviors* in two age groups of children (3 to 4 years old and 8 to 9 years old) from the general population. This study was designed in part to replicate the principle investigators' previous findings for 3 to 4 year old children and to extend those findings to test whether effects could be identified in 8 to 9 year old children from the general population. The study design was a double-blind, placebo-controlled, cross-over challenge with each treatment administered in fruit juice drinks daily for 1 week. Two mixes of artificial food colors and sodium benzoate were tested using several measures of *hyperactivity profile behaviors*, including parent ratings, teacher ratings and classroom observations, as well as computerized continuous performance testing in the older children. To measure individual differences in *hyperactivity profile behaviors* using these different sources of behavior measures, a Global Hyperactivity Aggregate (GHA) score was calculated as the unweighted composite aggregate of the standardized weekly parent rating, teacher rating, classroom observation, and continuous performance task (CPT) z-scores. A high GHA level indicated more *hyperactivity profile behaviors*.

Based on the analysis of the *whole sample* of children (considered the primary analysis), the findings from the primary study were that challenge with Mix A of artificial food colors and benzoate preservatives elicited statistically significant increases in GHA levels (greater *hyperactivity profile behaviors*) of 3/4 year old children and that challenge with Mix B elicited statistically significant increases in GHA levels (greater *hyperactivity profile behaviors*) of 8/9 year old children. Additional analyses were conducted to assess the effects of both challenge mixes on each individual (disaggregated) component behavior measure. While the latter analyses indicated that the parental rating was the major contributor for the primary effects of Mix A on GHA levels in 3/4 year old children, the parental rating measure alone did not show a statistically significant effect for Mix A challenge. In the 8/9 year old children the parental rating and continuous performance task (CPT) scores were the major contributors for the primary effects of Mix B on GHA levels but only the parental rating showed a statistically significant treatment effect for Mix B. In neither age group of children were there any significant treatment related changes detected by teacher ratings or classroom observations. The principle investigators' previous study with 3 year old children found a significant behavioral effect of Mix A based specifically on parental ratings. This previously reported effect of Mix A, based specifically on parental ratings was not replicated in the present study in the 3/4 year old children, although the Mix A challenge did elicit a statistically significant increase in the overall GHA levels (aggregate *hyperactivity profile behaviors*).

Overall, the primary study findings are suggestive of low level behavioral effects of a one week exposure to AFCAs on behavior in 3/4 year old children (Mix A) and 8/9 year old children (Mix B), based solely on parental ratings. However, due to the absence of confirmation of the parentally identified treatment effects by any other behavior measures together with the concerns about the data analyses and various procedural weaknesses of this study, it is the reviewers' opinion that there is questionable confidence in the reliability and biological relevance of the primary findings from this study. One particular procedural weakness relevant to regulatory application was the use of chemical mixtures as challenge materials which basically precludes identifying which specific compound(s) within the mixtures might be responsible for any treatment related effects. Consequently, there would be little, if any, utility of these findings to assess risk or to support regulatory decisions for specific compounds.

II. Introduction – Study Objectives

The primary hypothesis tested in the present study was that mixtures of certain artificial food colors and sodium benzoate (AFCA) increase the mean level of *hyperactivity profile behaviors* in two age groups of children (3 to 4 years old and 8 to 9 years old) from the general population. The study design was a double-blind, placebo-controlled, cross-over challenge with each treatment administered in fruit juice drinks daily for 1 week. Two mixes of artificial food colors and sodium benzoate were tested using several measures of *hyperactivity profile behaviors*, including parent ratings, teacher ratings and classroom observations, as well as computerized continuous performance testing in the older children. The more salient correlative secondary questions addressed in this study included:

- whether concordant treatment related effects are seen in teacher ratings, direct observations and test performance, as well as in parental ratings;
- whether behavioral and metabolic changes in children are apparent following acute challenge with a single dose of a food color/sodium benzoate mixture; and
- whether selected genetic differences (polymorphisms) modulate the behavioral effects of AFCA treatment.

III. Study Design/Methods

A. Primary Study

1. Participants

Two age groups of children in Southampton, UK were recruited for participation in this study, one group of 3-4 year olds (3/4YO) and one group of 8-9 year olds (8/9YO). The study sample for the 3/4YO group (n=153 with 79 males/74 females) was drawn from a general population of children registered in ‘early-years settings’ (EYS) (nurseries, day nurseries, preschool groups, playgroups). The sample for the 8/9YO group (n=144 with 75 males/69 females) was drawn from children attending primary and junior schools. Schools (nine participating) and EYS (26 participating) were selected to reflect the full range of socio-economic background of children in the area. Behavioral screening data from a hyperactivity questionnaire (ADHD Rating Scale-IV/Teacher version) rated prior to the start of the study by teachers for all 3-year-old and 8/9-year old children in participating schools and EYS to indicate the frequency of inattention and hyperactive behaviors over the previous 6 month period also showed that behavioral ratings for the study samples are representative of those for children of the same age in the participating schools and EYS [see *Reviewer Comment 1*, below]. At an initial home visit, written informed consent was obtained from parents who agreed to participate in the study. During the home visits, a report was obtained of each child’s pre-study diet based on 24-hour recall by the parents to assess the levels of foods containing additives consumed by the children in the previous 24 hour period. Prior to the start of the study the participating parents completed a behavioral questionnaire (ADHD Rating Scale-IV/Home-Parent version) to indicate the frequency of occurrence of inattention and hyperactive behaviors over the past 6 months [see *Reviewer Comment 2*, below].

2. Primary Study Design

The primary study design and challenge protocols were similar for both age groups of children (3/4YO and 8/9YO). Testing of the two age groups was conducted

consecutively with the 8/9YO group tested first. The experimental design was a double-blind within-subject crossover with two active treatments (Mix A and Mix B) and a placebo administered daily in a fruit-juice drink.

3. Challenge Treatments

The compositions of the two active Mixes for each age group are shown below in Table 1.

TABLE 1. Composition of Challenge Material

	Mix A		Mix B	
3/4 Year Olds	Sunset yellow	5.0 mg	Sunset yellow	7.5 mg
	Carmoisine	2.5 mg	Carmoisine	7.5 mg
	Tartrazine	7.5 mg	Quinoline yellow	7.5 mg
	Ponceau 4R	5.0 mg	Allura red AC	7.5 mg
	Total Colors	20 mg	Total Colors	30 mg
	Sodium benzoate	45.0 mg	Sodium benzoate	45.0 mg
8/9 Year Olds	Sunset yellow	6.25 mg	Sunset yellow	15.6 mg
	Carmoisine	3.12 mg	Carmoisine	15.6 mg
	Tartrazine	9.36 mg	Quinoline yellow	15.6 mg
	Ponceau 4R	6.25 mg	Allura red AC	15.6 mg
	Total Colors	24.98 mg	Total Colors	62.4 mg
	Sodium benzoate	45.0 mg	Sodium benzoate	45.0 mg

For the 3/4YO children the two active mixes (A and B) differed from each other in quantity of color additives and the specific color additives used. Mix A was similar to the active challenge used in the earlier study of 3-year old children by Bateman et al (2004) and Mix B was selected to represent the current average daily consumption of food additives by 3-year olds in the UK. Similarly, for the 8/9YO children the two active mixes (A and B) differed from each other in quantity of color additives and the specific color additives used. Mixes A and B used for the 8/9YO children had higher levels of color additives than the mixes used for the 3/4YO children, in order to account for the increased amount of food typically consumed by older children. However, both Mixes A and B in both the 3/4YO and 8/9YO groups included the same amount of sodium benzoate, a food preservative (see *Reviewer Comment 3*, below).

The placebo drink consisted of a mixture of fruit juices. The specific composition of the placebo drink and the variety of fruit juices used were not described [see *Review Comment 4*, below]. The only difference in the composition of the placebo and active mixes was the presence of the AFCA in the active mix with some variation in the proportions of the fruit juices to ensure matching color and taste for the placebo and active drinks. A masked testing by two independent panels of 20 young adults showed that the active and placebo juice drinks could not be differentiated based on look and taste. During the course of the study, the study administrator assigned the

challenge sequence (Mix A, Mix B and placebo) for each child. The child's family and the research team were masked to the challenge sequence. Identical sealed bottles of juice drinks were delivered to homes every week by the research team [see *Reviewer Comment 5*, below]. The juice was kept in a refrigerator and consumed at home either prior to the child's session in the EYS/school or after return from the EYS/school; the 8/9YO children consumed the juice mainly after returning from school (p. 168) [see *Review Comment 6*, below]. Any bottles with unconsumed juice were returned to the study office. Parents completed a daily diary of juice consumption and monitored compliance with the diet by recording dietary infractions ('mistake events'), when a child consumed a portion of food containing the artificial colors or sodium benzoate.

4. Challenge Protocol

After one week on the normal diet (week 0: baseline), the challenge protocol was conducted over a period of six weeks throughout which the artificial colors and sodium benzoate, to be used in the active challenges, were removed from the children's diet. During the first week of additive (colors/benzoate) withdrawal from the diet, all children received placebo drinks (week 1: withdrawal with placebo); during weeks 2, 4, and 6, each child was scheduled to a randomized set of two active challenge weeks and one placebo challenge week (week 2, 4 or 6: challenge) in which the children were given the appropriate challenge drink daily for seven days; and in weeks 3 and 5 all children received placebo drinks (week 3 or 5: washout with placebo). The 3/4YO children were given daily drinks of 300 mL/day and the 8/9YO children were given daily drinks of 625 mL/day. At the beginning of the study, the study administrator assigned each child using a random number generator to one of six possible sequences of receiving the placebo, active Mix A, or active Mix B challenges across weeks 2, 4, and 6.

5. Behavioral Measures

Behavioral Screening: As noted previously, behavioral screening, using the ADHD Rating Scale-IV questionnaires (teacher and parent versions), was conducted one time prior to Baseline: Week 0 by teachers for all children of the target age within participating schools and EYS (early year settings) and by participating parents to indicate frequency of specific inattention and hyperactive behaviors over the previous 6 months [see *Reviewer Comments 1 and 2*, below].

Weekly Behavior Measures: During the primary study, three measures of behavior were used to assess treatment effects for the 3/4YO children, with an additional fourth measure for the 8/9YO children. (1) Teacher ratings – the abbreviated ADHD Rating Scale-IV(teacher version) was completed by teachers once each week to indicate the frequency of inattention and hyperactive behaviors displayed over the past week for each week of the study (Week 0: baseline to Week 6). (2) Parent ratings – Parents of 3/4YO children used the abbreviated Weiss-Werry-Peters (WWP) hyperactivity scale and parents of 8/9YO children used an abbreviated ADHD Rating Scale-IV (parent version) to rate changes in their children's behavior over the previous week for each week of the study (Week 0: baseline to Week 6) [see *Reviewer comment 7*, below]. (3) Classroom Observation Code (COC) – The COC assesses the occurrence of mutually exclusive behaviors during structured didactic teaching and during periods of independent work under teacher supervision. In developing this measure,

behaviors had been selected to indicate components of ADHD that are shown in the classroom; for 3/4YO children, who in the UK have little structured didactic teaching and are not normally given “tasks” but allowed to choose a range of activities, the observation code was adjusted accordingly. Each child was observed by an independent observer (psychology graduate) for a total of 24 minutes each week of the study in three 8-minute observation sessions from which a total mean weekly score was derived [see *Reviewer comments 8 and 9*, below]. (4) Conners’ Continuous Performance Test II (CPTII) - A fourth behavioral measure for 8/9YO children was the CPTII, a computer based test of 14 minutes duration using response to visual stimuli to assess attention and the response inhibition component of executive control brain function. The CPT was administered weekly to the 8/9YO children only. The subject is presented with 18 blocks of 20 trials each (total 360 trials) and tested with three different inter-stimulus intervals (ISI) of 1, 2 or 4 seconds. This CPT is a “not X” task, requiring the subject to press a computer key immediately in response to all letter presentations other than the letter “X”. Four test measures (standard error of reaction time, % commission errors, signal detection index - d' , and response bias - β) were used to derive a weekly CPT aggregate score. These measures have been shown to be highly correlated with the parent ADHD rating scale measure.

6. Data Analysis Methods

Global Hyperactivity Aggregate (GHA): A Global Hyperactivity Aggregate (GHA) score was developed to measure individual differences in hyperactivity (*hyperactivity profile behaviors*) using different sources of behavior measures (i.e., teacher ratings, parent ratings, classroom observations, and a computerized CPT test). To calculate the GHA, weekly scores from the Parent, Teacher, COC and CPT measures for each child were standardized to time 0 at Baseline (T0; aka Week 0) for the same measure, as:

weekly standardized (z) aggregate score = (score X – mean X at T0) / SD at T0

The GHA was then calculated as the unweighted composite aggregate of the weekly Parent, Teacher, COC, and CPT z-scores. The GHA was calculated only when at least 3 of the different behavior z-scores were present for any week. The absolute value of the GHA score indicates the relative change in hyperactivity compared to baseline (the higher or more positive the number, the greater the level of hyperactivity behavior compared with baseline and the lower or more negative the number, the lower the levels of hyperactivity relative to baseline).

Statistical Analysis: Although the study designs for the two age groups (3/4YO and 8/9YO) were similar, the difference in composition of the GHA (the additional CPT behavior measure was included in testing of 8/9YO children only), and in the dose and composition of the AFCA mix used, precluded joint analysis of the data from the two age groups. Therefore, for analysis purposes the studies for the two age groups were treated as parallel but independent.

Linear mixed-model methods in Statistical Package for the Social Sciences (SPSS) were used to analyze data. Two models were tested separately for each age group for the effects of treatment on GHA in challenge weeks. Model 1 used the challenge type alone as a fixed effect testing for Mix A versus placebo and Mix B versus placebo. In Model 2, in addition to challenge type, the effects of the following potential confounding factors were controlled: week during study, sex, GHA in baseline week

(Week 0), number of additives in pretrial diet, maternal educational level, and social class [see *Reviewer Comment 10*, below]. The study was powered to detect differences between the active and placebo periods and, accordingly, the effects of Mix A and Mix B were compared with that of placebo.

The analyses were replicated for three subgroups of the study population in each age group: the *whole sample*, a high consumption sub-sample group ($\geq 85\%$ consumption of drinks in any challenge week), and a *complete case* sub-sample (consuming $>85\%$ of drinks in all challenge weeks and no missing GHA score). The latter two sub-sample groups were included to determine whether non-compliance (not consuming all of the scheduled drinks) and the method of handling missing data affected the pattern of results.

B. Additional Experimental Design Procedures to Address Salient Correlative Secondary Research Questions

1. Effects of the AFCA challenge mixes on component behavior measures of GHA
The present study was designed with a global aggregated measure of hyperactivity (GHA), combining the behavior measures (teacher ratings, parent ratings, classroom observations, and a computerized test), as the primary outcome measure. Treatment effects were calculated as changes in GHA in response to the challenge mixtures compared with placebo. In a previous study of the behavioral effects of AFCAs in 3-year old children (using a challenge equivalent to Mix A in the present study), uncertainty in the interpretation of the results occurred due to the finding that significant effects of the Mix A challenge were detected only by parental reports and were not confirmed by clinical behavioral testing (Bateman et al, 2004). In an effort to determine whether the previous findings were replicable and to extend those findings, additional analyses of the behavior data for the 3/4YO and 8/9YO children in the present study were conducted to address the correlative secondary research question of whether challenge related effects are seen across the different sources of behavior measures in addition to the parent ratings. For this purpose, analyses of the behavioral response of additive challenge versus placebo for each individual behavior measure (*disaggregated measures* of teacher ratings, parent ratings, classroom/playroom observation scores, and CPT) were performed and presented in the final report as a secondary outcome measure.
2. Behavioral effects and metabolic mediators following acute challenge with AFCA
A second phase of this study was designed as a “proof of principle” component to address the correlative secondary question of whether it is possible to demonstrate short term changes in behavior immediately after single dose acute challenge with a mix of AFCAs and to explore their relationship to metabolic factors that may mediate such responses. Participants in this study were enlisted from the 8/9YO boys who participated in the primary phase of the challenge study. Two groups of 15 boys each, who did or did not exhibit a behavioral response to Mix B in the primary challenge study, were identified. Using each child’s GHA scores from the primary study, a GHA difference score was calculated by subtracting the Placebo score from the Mix B score. Higher difference scores reflected a more negative behavioral response to Mix B. The GHA difference scores were then ranked and children with scores $\geq 75^{\text{th}}$ percentile were classed as “*responders*” and those with scores $\leq 25^{\text{th}}$

percentile as “*non-responders*”. Parents of children were asked to resume the reduced additive diet that was followed in the primary phase of this study for a period of 24 hours prior to initiation of the acute challenge phase of study. The Acute Challenge was conducted at the Southampton School of Psychology in two 2.5 hour sessions approximately one week apart. In each session children were given in a random sequence a single dose of either the active challenge (Mix B with amounts of additives equivalent to the daily total administered in the primary study) or the placebo challenge, each administered by capsule. The children’s response inhibition and attention were measured using the Conners’ Continuous Performance Test II (CPT) twice, approximately 30 minutes prior to and 30 minutes after challenge. While completing the CPT, children’s behavior was also monitored by independent observers using the seven item Hillside Behavior Rating Scale. In addition to the behavioral measures, urine samples were collected to test for histamine which is thought to mediate a pharmacological effect of additives in food, and saliva samples were collected for assaying levels of tryptase activity, a possible marker of inflammatory processes. Urine samples were collected prior to the laboratory visit and again at 15 minutes and 50 minutes after challenge. Saliva samples were collected at approximately 50 minutes and 15 minutes pre-challenge and again at approximately 15 minutes and 50 minutes post-challenge. For data analysis, scores for the CPT and the observational ratings were converted to z-scores and aggregated to produce a Hyperactivity Index (HI) score for the pre-challenge and the post-challenge periods [see *Reviewer Comment 11*, below]. The pre-challenge HI was subtracted from the post-challenge HI to produce an HI difference score for statistical analysis. Higher difference scores reflected more negative behavior in response to the Mix B challenge. Mixed Model methods were used to analyze the data [see *Reviewer Comment 12*, below]. Only the results of the behavioral analyses were presented in the final report. The results relating to the metabolic factors were not complete at the time the technical report was prepared and are to be presented at a later date.

3. Modulation of AFCA behavioral effects by genetic polymorphisms

To address the secondary question of whether genetic differences make individual children more or less sensitive to the AFCA (Mix A and Mix B) treatments, selected genetic polymorphisms were analyzed using DNA from cheek cells for all 3/4YO and 8/9YO children of both sexes in the main study [see *Reviewer Comment 13*, below]. Single nucleotide polymorphisms (SNPs) were selected from the dopamine and adrenergic neurotransmitter system since these have previously been implicated in ADHD. Since there is also a suggestion that histamine may be involved in the effects of AFCAs, genetic polymorphisms from this system were also included. Results were presented in the final report for two SNPs in the histamine N-methyltransferase gene (HNMT Thr105Ile) and (HNMT T939C), one SNP in the dopamine related catechol-o-methyltransferase gene (COMT Val108Met), and one SNP in the adrenergic neurotransmitter receptor alpha 2A gene (ADRA2A C1292G). The Mixed Models analyses used to determine whether these genotypes modulated the effects of the AFCAs (Mixes A and B) were limited to those children consuming an adequate amount of the challenge ($\geq 85\%$ consumption sub-group), since the aim of this analysis was not to establish the impact of the individual additives per se (where the intention to treat based on the *whole sample* is the focus). Thus, the analyses focused on the interactions between genotype and effects of Mix A and Mix B in the $\geq 85\%$

consumption sub-sample of 3/4YO and 8/9YO children.

C. Reviewer Comments Regarding Study Design/Methods

1. **Questionable accuracy of teacher pre-study behavioral ratings recalling behavior over a period of 6 months**

The teachers completed behavioral questionnaires prior to the start of the study to indicate frequency of certain behaviors for individual children over the previous 6 months. The accuracy of ratings based on a recall of behaviors over a period of 6 months is highly questionable, particularly for teachers rating each of multiple children in a classroom or playroom setting.

2. **Questionable use and accuracy of pre-study behavioral ratings by participating parents**

As noted above for the teacher screening data, the accuracy of pre-study ratings based on a recall of behaviors over a past 6 month period would be questionable. Also, the information derived from these parental ratings did not appear to be used for any analyses in the study itself.

3. **Disproportionate increases were made in levels of artificial colors in the mixes (A and B) used for the 8/9YO children relative to the mixes (A and B) used for the 3/4YO children, and no adjustments were made to levels of sodium benzoate in either Mix A or B for either age group of children**

As stated in the technical final report (p. 10), the levels of colors in Mix A used for the 8/9YO children were increased by 1.25 times the levels of colors in Mix A used for the 3/4YO children. However, Mix B levels of colors for the 8/9YO children were disproportionately increased by 2.08 times the levels in Mix B used for the 3/4YO children. Although the rationale for adjusting levels of colors in the mixes for both age groups of children included attempts either to reflect the current average daily consumption of food additives by 3-year-olds in the UK (Mix B) or to account for the increased amount of food typically consumed by children in the 8 to 9 year old range, there was no explanation why comparable adjustments in sodium benzoate levels were not made across mixes or across age groups. The differences in levels and types and quantities of colors and the lack of adjusted levels of sodium benzoate across mixes (A and B) and particularly across age groups essentially precludes any reliable dose response evaluations or comparative assessments of treatment effects between Mix A and Mix B within or between age groups of children.

4. **Composition of placebo is unclear**

Other than the placebo drink being made of various proportions of fruit juices, no specific information was provided about the composition of the placebo (in Annex 2, p. 121 of final report one mention was made of two 8/9YO children excluded from the study due to allergic reaction to *blackcurrant juice*). The placebo is a critical component of this challenge study and it would be important to have information about the types of fruit juice used (commercial or fresh). If commercial, what brands were used and did they contain any types of additives? If fresh, how was the juice made and when was it prepared relative to use in the study? Such information about the placebo source and composition are very important in helping to assess the outcome information from this study.

5. Parents were not asked about the blinding

Although it was determined that the active and placebo juice drinks could not be differentiated based on look and taste and the parents were blind to which challenge drink (active or placebo) was being delivered in any particular week, the investigators did not specifically ask the parents whether they could tell when the active challenge was being given at any time during the study.

6. Timing of juice consumption at home

While all children in both age groups consumed the juice drinks at home, they consumed the drinks either prior to or after their session at EYS or school (8/9YO children consumed their juice mainly after returning from school). However, there is no information provided as to the percent of children that consumed the juice drinks before going to EYS or school or going for classroom observation or CPT session. This information would be important for interpreting detection of treatment effects relative to latency to onset of effects after dosing.

7. Precision of recall data in conduct of weekly parent/teacher ratings and availability of standard ratings

The parent and teacher weekly ratings during the study proper were based on recall of behaviors for individual children over the previous week. Recalling behaviors over a previous seven day period, although acceptable, will invariably be less precise than daily ratings. However, daily ratings would have presented logistical difficulties. The investigators did not provide comparative standard teacher/parent behavioral ratings for ADHD and non-ADHD children, with which to gauge the significance of the ratings for the children in this study.

8. Reliability of modified Classroom Observation Code (COC) used in 3/4YO children is unclear

On page 131 of the final report, the COC used for the 8/9YO children was noted as having adequate inter-observer reliability and ability to discriminate between hyperactive and non-hyperactive children (no definition or criteria provided for designation of ADHD). Since a modified version of the COC was used for the 3/4YO children, it is not clear whether this modified COC was equally reliable. The investigators did not provide comparative standard COC ratings for ADHD and non-ADHD children, with which to gauge the significance of the COC for the children in this study.

9. Questionable adequacy of duration used for classroom observation

Each child's weekly classroom observation score was based on only three 8-minute observation periods (total of 24 minutes of observation per week). It is questionable whether observation for 8 minutes, three times a week is an adequate sampling frequency or observation duration to provide a reliable measure of representative classroom behavior over an entire week.

10. Model 1 is not useable for analysis of data

Since statistical Model 1 does not adjust for the potential confounds in the study, the analyses using Model 1 do not provide reliable information about treatment related effects and are of no apparent value in the final evaluation of findings.

11. In the “acute challenge phase” of this study the disaggregated behavior measures comprising the Hyperactivity Index (HI) were not analyzed separately

The component behavior measures, CPT and independent behavior ratings were aggregated to form the Hyperactivity Index (HI). However, these component behavior measures were not analyzed separately (disaggregated) to determine whether changes in CPT performance or the behavior rating contributed more, less or equivalently to treatment related changes in the overall aggregate HI scores.

12. The Mixed Model analyses of the “acute challenge” data did not appear to control for potential confounding factors

In describing the acute challenge findings, the final technical report does not indicate that any potential confounding factors were either identified or controlled for in the Mixed Model methods used to analyze the data. In the absence of controlling for confounding variables there is low confidence in the reliability of the statistical analyses of the acute challenge study findings.

13. No description of materials or methods was provided for the genotyping analyses

The investigators did not provide any details in the final report about the manner in which DNA samples were taken, at what period during the main study samples were taken, how the samples were maintained, or how the samples were analyzed for SNPs. Without such methodological information it was difficult to assess the adequacy of the procedures used in this analysis or the reliability of the data collected.

IV. Study Investigators’ Reported Findings

A. Reported Study Findings Relative To Primary Research Question: Do Mixtures of Certain Artificial Food Colors and Sodium Benzoate Increase the Mean Level of Hyperactive Behavior (Hyperactivity Profile Behaviors) In Children From the General Population?

1. Study Sample and Background Characteristics

Table 2 shows the numbers of 3/4YO and 8/9YO subjects in the *whole sample*, in a sub-group of children who had *consumed* $\geq 85\%$ juice drinks in any challenge week over the period of the study, and in a smaller sub-group of children who consumed $\geq 85\%$ juice drinks and also had no missing behavior data (*complete case* group).

TABLE 2. Numbers of 3/4-Year Old and 8/9-Year Old Children in Study

	Children in Whole Sample	Children Consuming $\geq 85\%$ of Challenge Drinks	Children in Complete Case ($\geq 85\%$ Consumption & No Missing Data)
	N	N*	N*
3/4-Year Olds	153 (79 m/ 74 f)	133	73
8/9-Year Olds	144 (75 m/ 69 f)	119	91

* The sex distributions in the $\geq 85\%$ Consumption group and the Complete Case group were not specified.

The analysis of background characteristics, including race, marital status, social class/employment of father and of mother, and mother’s education level, revealed no significant differences between these three groups or between the groups of children assigned to receive the challenge drinks in different orders over the course of the study.

Based on pre-study teacher’s behavioral ratings (ADHD Rating Scale-IV/Teacher version), the sample populations of male and female subjects in both the 3/4YO and 8/9YO age groups were representative of the general populations of age-matched children within the early year settings and schools in terms of the Teacher scores for inattention, hyperactive and total behaviors. As expected, significant gender differences were found in each age group for both the population ($p < 0.001$) and study sample ($p < 0.001$) with boys having higher behavioral scores than girls.

2. Juice consumption and dietary infractions

Juice consumption over the period of the study remained at what was considered an acceptable level for the majority of children in both age groups. Of the children who completed the study 93% of the 3/4YO children and 85% of the 8/9YO children consumed more than two thirds of all drinks and 75% of both age groups consumed $\geq 85\%$ (at least 6 out of 7 daily drinks per week). Dietary infractions during the study were considered acceptably low for both age groups with approximately 30% having 0 infractions and 17% having more than 4 infractions. Rates were comparable during active and placebo weeks.

3. Effects of Challenge

The unadjusted mean GHA scores for 3/4YO and 8/9YO children by challenge type are presented in Table 3.

TABLE 3. Unadjusted Mean GHA Scores* for 3/4YO and 8/9YO Children by Challenge Type

		<u>Mix A</u> n	<u>Mix A</u> Mean (SD)	<u>Mix B</u> n	<u>Mix B</u> Mean (SD)	<u>Placebo</u> n	<u>Placebo</u> Mean (SD)
3/4 Year Old Children	Whole Sample	131	- 0.11 (1.03)	134	- 0.14 (1.03)	129	- 0.32 (1.11)
	$\geq 85\%$ Consumption	104	- 0.11 (1.03)	108	- 0.15 (1.07)	99	- 0.39 (1.07)
	Complete Case	73	- 0.14 (1.04)	73	- 0.26 (1.05)	73	- 0.44 (0.98)
8/9 Year Old Children	Whole Sample	132	0.25 (0.97)	133	0.33 (1.10)	127	0.19 (1.03)
	$\geq 85\%$ Consumption	104	0.26 (0.93)	112	0.32 (1.09)	103	0.19 (1.04)
	Complete Case	91	0.27 (0.92)	91	0.35 (1.08)	91	0.19 (1.06)

* A numerically higher GHA score indicates a greater level of hyperactivity (*hyperactivity profile behaviors*)

It should be noted that the GHA scores (which indicate level of behavior relative to T0 baseline prior to removal of AFCA from the diet and prior to any challenge drinks being administered) for the 3/4YO children in all groups were generally below the baseline level (i.e., less *hyperactive profile behaviors* relative to the T0 baseline). This was interpreted to mean that the effects of reduced hyperactive profile behaviors from withdrawal of AFCA from the diet were not being counteracted by the effects of subsequent AFCA challenges. For the 8/9YO children, however, the GHA scores overall tended to increase above the baseline level (i.e., more *hyperactive profile*

behaviors relative to baseline). This was attributed to an overall decreasing level of CPT performance across the study period for this age group of children, which contributed to an overall increase in their GHA scores (CPT was not administered to the 3/4YO children).

In identifying potential moderating or confounding effects of variables on the behavioral response to challenge, preliminary Mixed Models analyses showed that for both age groups of children the baseline (T0) GHA behavior score was related to the challenge GHA at all subsequent time points and that gender and pre-study diet (3/4YO children only) were individually related to baseline GHA but in no case was there a significant interaction between these or other factors and challenge type (i.e., AFCA or placebo). There was no effect of carryover from challenges in previous weeks on behavior during subsequent challenge weeks. While the 3/4YO children showed no effect of time (week) on the GHA score, for the 8/9YO children there was an effect with GHA increasing across weeks during the study. An examination of the component behavior measures of the GHA indicated that this was due to a gradual worsening of the children's scores on the CPT over the weeks of the study. This was attributed to the children becoming less motivated to perform this intrinsically boring task with repeated testing. The declining CPT performance week by week resulted in increasing GHA scores for the 8/9YO children [see *Reviewer Comment 1*, below].

The final assessment of significance of the effects of challenge (Mix A or Mix B versus placebo) on GHA is based on a Mixed Model analysis with potential confounding factors controlled, including the effects of week during study, gender, GHA in baseline week (T0), number of additives in pre-test diet, maternal education level and social class. The results from the analyses of estimate effect sizes (Mix A versus placebo and Mix B versus placebo) for both 3/4YO and 8/9YO groups of children are shown in Table 4, separately for the full sample (primary outcome), and *post hoc* analyses of two sub-samples one with a high $\geq 85\%$ consumption of juice drinks in any challenge week and the other, the *complete case* sub-sample, with $\geq 85\%$ consumption in all challenge weeks and no missing behavioral GHA scores [see *Reviewer comment 2*, below].

TABLE 4. General Linear Mixed Models estimates of GHA effect size for challenge Mix A versus Placebo and challenge Mix B versus Placebo

[*Whole Sample*, $>85\%$ Consumption, and *Complete Case* Groups of 3/4 Year Old and 8/9 Year Old Children with all potential confounding factors controlled]

Subjects	Challenge Type	Estimate Effect Size (95% CI)					
		N	Whole Sample	N	$\geq 85\%$ Consumption	N	Complete Case
3/4 Year Old Children	Mix A vs Placebo	140	0.20 (0.01 to 0.39)*	130	0.28 (0.05 to 0.51)*	73	0.32 (0.05 to 0.60)*
	Mix B vs Placebo	140	0.17 (- 0.03 to 0.36)	130	0.19 (-0.04 to 0.41)	73	0.21 (- 0.06 to 0.48)
8/9 Year Old Children	Mix A vs Placebo	136	0.08 (- 0.02 to 0.17)	119	0.09 (- 0.01 to 0.19)	91	0.12 (0.02 to 0.23)*
	Mix B vs Placebo	136	0.12 (0.03 to 0.22)*	119	0.15 (0.05 to 0.25)**	91	0.17 (0.07 to 0.28)**

* p<0.05, ** p<0.01

All significant treatment related effects on behavior relative to placebo in both age groups of children (Table 4) were in the direction of increased hyperactivity profile behaviors. For the 3/4YO children, the primary analysis of the data using the *whole sample* showed a statistically significant effect of Mix A on GHA compared to placebo ($p < 0.05$) but no significant effects of the Mix B challenge. When the analysis was limited either to those children with $\geq 85\%$ juice consumption or to the *complete case* sub-samples, the adverse effect of Mix A on behavior remained statistically significant ($p < 0.05$) and Mix B still had no significant effects.

For the 8/9YO children, the primary analysis of the *whole sample* showed a statistically significant effect of Mix B on GHA compared to placebo ($p < 0.05$) but not Mix A. The same pattern of effect occurred when the analysis used the data from the $\geq 85\%$ consumption sub-group, that is the effect of Mix B remained significant ($p < 0.01$) but Mix A had no significant effects. When the analysis was limited to the *complete case* sub-sample with $\geq 85\%$ consumption and no missing GHA data, the effects of both Mix A and Mix B were statistically significant ($p < 0.05$ and $p < 0.01$, respectively) [see *Reviewer comment 3*, below].

B. Reviewer Comments Regarding Reported Study Findings for Primary Research Question

1. Limited information regarding CPT performance decrements of 8/9YO children over study period

CPT performance of the 8/9 year old children declined over the course of the study. Although little detailed information was presented, the study investigators concluded that this decline was attributable to the children becoming bored with the repeated conduct of this inherently tedious task. Sufficient additional information should have been provided to support this conclusion.

2. Problematic aspects of the data analyses

The analyses in this study were replicated in each age group for the *full sample* of subjects, for a high consumption group ($\geq 85\%$ consumption of juice drinks in any challenge week), and for a *complete case group* ($\geq 85\%$ consumption in all challenge weeks and no missing GHAs). The rationale for including the latter two groups was to determine whether non-compliance (failure to consume challenge drinks) and the method of handling missing data affected the pattern of results. As shown in the findings from the analyses of the overall GHA scores, there was some variability in the occurrence of statistically significant effects across subgroups. Specifically, when the *whole sample* or $\geq 85\%$ consumption groups of 8/9YO children were used for analysis, only Mix B had a significant effect. However, when the *complete case* group of 8/9YO children was used, both Mix A and Mix B had significant effects on GHA scores. Additional more notable examples of variability across subgroups occurred in the analyses of the change in GHA scores for the component (disaggregated) behavior measures (shown in Section IV, C below). Yet, there was little, if any, qualification accompanying the description of the data analyses or the results using the whole sample or the two sub-samples of subjects. This resulted in confusion as to which analyses provided the primary study results. However, the principle investigators eventually stated in the Discussion of the Primary Research Question Findings (p. 26) that the use of the “*whole sample*” of the 3/4YO and of the 8/9YO children is considered to be “the primary analysis of the data on an intention

to treat basis”. While the variability in significant treatment effects using the different subgroups of subjects for analysis may indicate that non-compliance in consumption of challenge drinks or the method of handling missing data affected the pattern of results, the investigators do not discuss the impact of this problematic aspect of the data analyses on interpretation of the study findings.

3. **There is no clarification of the behavioral changes in this study as necessarily “adverse” (p.22) or “deleterious” (p.26)**

There was no information provided in this study to suggest that the changes in behavior based on parental reports were adverse, detrimental or maladaptive in any demonstrable manner. The children’s behavior under challenge conditions appeared to be within the range of behavioral levels exhibited by the general population of age-matched children. The significant statistical findings should be presented simply as “effects” and not “adverse” or “deleterious” effects, without further clarification.

C. Reported Study Findings Regarding Secondary Research Questions

1. (Secondary Research Question) Effects of the AFCA challenge mixes on component behavior measures of GHA

To gauge the extent to which the individual behavior components of the overall GHA may have contributed to the effects of AFCA treatment on children’s behavior, additional statistical analyses were conducted as a secondary outcome measure. These analyses considered the effect of challenge on the disaggregated standardized GHA scores for each of the component behavior measures; teacher ratings, parent ratings, classroom observation scores, and CPT. The results of these analyses are presented in Table 5.

In presenting the results of analysis of the effects of challenge on the disaggregated GHA scores for individual behavior measures, the principle investigators state that any single indicator is likely to be relatively less reliable compared to the overall composite GHA. The consequent increased measurement error in the analysis of individual behavior measures makes it less likely that a significant effect will be detected. For this reason the principle investigators concluded that the results for the disaggregated behavior measures are most appropriately discussed in terms of the effect sizes of Mix A versus placebo and Mix B versus placebo [see *Reviewer Comment 2(a)*, below]. With specific reference to the 3/4 YO children (p.104), the investigators suggested that the findings should be viewed in the context of the additive mixes being consumed at home and not at the early years setting (or in the classroom observation setting) and that any behavior score will be a function, among other things, not only of the amount of juice consumed but also the time and individual differences in absorption of additives.

TABLE 5. Effect of sizes of Mix A versus Placebo and Mix B versus Placebo for each disaggregated behavior measure

[*Whole Sample*, $\geq 85\%$ Consumption, and *Complete Case* Groups of 3/4 Year Old and 8/9 Year Old Children with all potential confounds controlled]

		Mix A vs Placebo (Estimate Effect Size)			Mix B vs Placebo (Estimate Effect Size)		
		Whole Sample	$\geq 85\%$ Consumption	Complete Case	Whole Sample	$\geq 85\%$ Consumption	Complete Case
3/4 Year Olds	Parent rating	0.33	0.49 ($p < 0.016$)	0.55 ($p < 0.027$)	0.27	0.36	0.37
	Teacher rating	0.01	0.03	0.09	0.06	0.08	0.10
	Classroom Observation	0.09	0.10	0.08	0.001	- 0.01	- 0.02
8/9 Year Olds	Parent rating	0.01	0.03	0.03	0.13 ($p = 0.031$)	0.13 ($p = 0.046$)	0.08
	Teacher rating	- 0.04	- 0.01	0.00	- 0.03	0.01	0.04
	Classroom Observation	0.02	0.08	0.04	0.01	0.05	0.07
	Continuous Performance Task (CPT)	0.10	0.08	0.18	0.19	0.20	0.31 ($p = 0.015$)

As presented in Table 5, the analyses of the challenge versus placebo effect sizes for the individual component measures of behavior were analyzed for *the whole sample*, for the subset *consuming* $\geq 85\%$ or more of the drinks, and for the *complete case* subsample of 3/4YO and 8/9YO children [see *Reviewer Comment 2(b)*, below]. Overall, the largest effects in the 3/4YO children were found for the parental ratings with both Mix A and B challenges. For the 8/9YO children the largest effects overall were found for the computerized test of attention (CPT) with both Mix A and B challenges, and moderate effects were found for parental ratings with Mix B challenge. In general, for both age groups a majority of the challenge versus placebo effect sizes for all behavior measures under both Mix A and Mix B challenge conditions were in the direction of increased hyperactivity (*hyperactivity profile behaviors*).

Specifically, for the *whole sample* of 3/4YO children the largest effects were those based on parental reports with both Mix A and Mix B challenges, neither of which was statistically significant. For the *whole sample* of 8/9YO children the largest effects were found for the computerized test of attention (CPT) under both Mix A and Mix B challenges, neither of which was statistically significant, but the moderate effects based on parental ratings were significant ($p < 0.04$) for Mix B but not Mix A. Notably, negligible non-significant effect levels were found for teacher ratings and classroom observations in the *whole sample* of both 3/4YO and 8/9YO children.

For the $\geq 85\%$ consumption and the *complete case* groups of 3/4YO children the largest effects were based on parental reports for both Mix A and Mix B, with only the parental report effect size for Mix A being statistically significant in both the $\geq 85\%$ consumption group ($p < 0.02$) and the *complete case* group ($p < 0.03$). For the 8/9YO children the largest effects were still found in the CPT under both Mix A and

Mix B challenges, with only the CPT effect size for Mix B in the *complete case* group being statistically significant ($p < 0.02$). In the analyses of the $\geq 85\%$ *consumption* and the *complete case* groups of 3/4YO and 8/9YO children, the only other statistically significant effect was the parental report effect for Mix B in the $\geq 85\%$ *consumption* group of 8/9YO children ($p < 0.05$). Negligible non-significant effect levels were still found for teacher ratings and classroom observation ratings using either the $\geq 85\%$ *consumption* or the *complete case* sub-samples of 3/4YO and 8/9YO children.

2. **Reviewer Comments: Effects of the AFCA challenge mixes on component behavior measures of GHA**

(a) **Complementary changes across disaggregated behavior measures can provide reliable confirmation of significant treatment effects.**

The principle investigators minimized the significance and utility of disaggregated measures by stating that they consider analysis of disaggregated measures to be less sensitive for finding a statistically significant effect and thereby less reliable than the global aggregate measure, GHA. They state that “Any single indicator is likely to be relatively less reliable compared to the aggregate measure. The consequent increased measurement error makes it less likely that a significant effect will be detected.” However, it should be noted that analyses of disaggregated behavior measures were sufficiently sensitive to provide a number of statistically significant findings. The statistical argument proposed by the principle investigators appears to support the use of artificially enhanced detection sensitivity by combining different types of behavioral measures (parent ratings, teacher ratings, classroom observation codes, and continuous performance testing), which effectively serves to lower the error term and enhance the likely occurrence of statistical significance. However, sensitivity (whether artificially increased or not) is not the only, nor is it arguably the most important, aspect of reliability in detecting a true treatment related effect. Replication of findings and concordance of changes across the various behavior measures, which serve to help confirm the occurrence of treatment effects, are also important elements of reliability. The present study focuses on “hyperactivity” as the target behavior. As used by the investigators in this study, the term “hyperactivity” indicates a behavioral profile characterized by overactive, impulsive and inattentive behavior. Finding complimentary statistically significant treatment related changes in the hyperactivity profile across several of the various behavior measures would seem to be more informative about the scope of effect and provide more confidence of confirmation of relevant treatment effects than a statistical finding with the enhanced sensitivity of a single aggregate score. No such complimentary treatment effects were found.

(b) **Use of the three sample groups of children for analysis of the component (disaggregated) behavior measures results in variable statistical outcomes but relevance of this to data interpretation are not discussed**

[See also *Reviewer Comment in section IV, B, 2* above regarding analysis of overall GHAs in main study]. As seen in the analyses of the treatment-related changes in the disaggregated GHAs for the component behavior measures, there is notable variability in the outcome of statistically significant effects depending upon whether the *whole sample*, the $\geq 85\%$ *consumption* sub-sample, or the *complete case* sub-sample of children was used for the analysis. For example, based on analysis of the component behaviors using the *whole sample* of children, there were no significant effects for the 3/4YO children, and for the 8/9YO children only the parent ratings were significantly affected with Mix B challenge. None of the teacher ratings or

classroom observation ratings (both age groups) or CPT scores (8/9YO) were significantly affected. However, when the analyses used the *complete case* group of children, parent ratings were significantly affected only for the 3/4YO children with Mix A, and CPT scores were significantly affected but only for the 8/9YO children with Mix B. The teacher and classroom observation ratings were still not significantly affected. Although this obvious variability in significant treatment effects across subgroups of subjects may indicate that non-compliance in consumption of challenge drinks or the method of handling missing data affected the pattern of results, the investigators do not discuss the impact of this problematic aspect of the data analyses on the confidence in the statistical significance of the study findings and their interpretation.

3. (Secondary Research Question) Behavioral effects and metabolic mediators following acute challenge with AFCA – Study Findings and Interpretation

The acute challenge component of the project explored the possibility of demonstrating short term changes in hyperactive behavior (*hyperactivity profile*) immediately post-challenge with Mix B for two groups of 8/9YO boys who did (n=15) or did not (n=15) respond to Mix B in the main challenge study. The results of the Mixed Models analyses with groups combined showed no statistically significant main effect of challenge with Mix B on the Hyperactivity Index (a composite aggregate of the CPT scores and the observational ratings) compared to Placebo (p=0.096), no significant difference between the Hyperactivity Index levels of responder and non-responder groups (p=0.134), and no significant Challenge x Group interaction on the response to Mix B challenge (p=0.072). Although all elements of the analyses were non-significant, there was a general trend towards increased *hyperactive profile* behavior in response to Mix B challenge, particularly by the group of “responders” compared to the “non-responders” [see *Reviewer Comment (a)*, below].

The principle investigators contended that the findings from the acute challenge study suggest that the 8/9 year old “responders” to challenge with Mix B exhibit effects within a short period of time after dosing (approximately one hour). The children in the main study consumed their challenge drinks at home. The 3/4 YO children consumed their drinks either prior to or upon returning from the early year setting but the 8/9 YO children usually consumed their drinks upon return from school (percentages were not reported). With an approximate one hour latency to onset of acute effects, the investigators suggested that this makes it likely that the treatment-induced behavior changes occurred in the home setting and this may be an explanation as to why the strongest effects were found for the parental ratings but not for the school or early years settings based measures [see *Reviewer Comment (b)*, below].

4. Reviewer Comments: Behavioral effects and metabolic mediators following acute challenge with AFCA – Study Findings

(a) Statistical analyses of data in the “acute challenge” component study did not control for confounding variables.

None of the confounding variables which were controlled in the data analyses of the primary study, such as CPT aggregate in baseline week, number of additives in pre-trial diet, maternal educational level and social class, appear to have been

incorporated into the Mixed Model methods used for analysis of the acute challenge data. Since potential confounds were not controlled in the data analysis, this calls into question the reliability of the findings from the acute challenge component study.

(b) Rationale for explaining why treatment effects in the main study were detected only in parental ratings was based on inaccurate interpretation of acute study findings.

Aside from the questionable reliability of the data analysis in the acute challenge study and the fact that no significant treatment effects were found, the trends in the acute challenge study (although not statistically significant) suggest that the onset of response following ingestion of Mix B challenge occurred within a short period of time (an hour). Since all of the children in the primary study consumed their challenge drinks at home (the 8/9 YO children usually consumed their challenge drinks after returning from school and the 3/4 YO children consumed their drinks either prior to or upon returning from their school setting), it seems likely that the parents in the home setting would have observed behavioral changes that might have occurred in the children. However, the acute challenge study was not designed to determine the duration or latency of treatment effects after challenge.

5. (Secondary Research Question) Modulation of AFCA behavioral effects by genetic polymorphisms – Study Findings

The results of the analyses examining the modulation of the effect of challenge by the children's genotype are shown in Table 6 for the 3/4YO children and Table 7 for the 8/9YO children. The analysis of the datasets for each genotype included an initial test for the main effect of challenge (Mix A and Mix B) on GHA levels, independent of whether the genotype was present or absent. In the 3/4YO children there was a significant main effect of challenge on GHA levels with both Mix A and Mix B in the analyses for HNMT Thr105Ile ($p=0.004$ and $p=0.02$, respectively) and HNMT T939C ($p=0.005$ and $p=0.036$, respectively), but there was no significant main effect with either challenge mix in the analysis for COMT Val108Met or ADRA2A C1291G. Comparably, in the 8/9YO children the main effect of challenge on GHA levels was significant with both Mix A and Mix B in the analysis for HNMT Thr105Ile ($p=0.046$ and $p=0.001$, respectively) and with Mix B (but not Mix A) in the analysis for ADRA2A C1291G ($p=0.036$). There was no main effect of challenge with either Mix A or Mix B in the analysis for the COMT Val108Met dataset. Although these results for main effect of challenge were not discussed to any great extent by the principle investigators in the final report, it seemed appropriate for purposes of completeness to describe these particular results in more detail in this review [see *Reviewer Comment (a)*, below].

TABLE 6. Mixed Models analyses of interactions between Genotype and Behavioral Responses (GHA) to Additive Mix A and Mix B versus Placebo in the $\geq 85\%$ consumption sub-sample of 3/4 YO Children with All Potential Confounds Controlled

		[Effect Size (p value)]			
		HNMT Thr105Ile	HNMT T939C	COMT Val108Met	ADRA2A C1291G
Additive Challenge	Mix A vs Placebo	0.39 (*p=0.004)	0.42 (*p=0.005)	0.39 (ns: p=0.087)	0.27 (ns: p=0.110)
	Mix B vs Placebo	0.30 (*p=0.020)	0.30 (*p=0.036)	0.11 (ns: p=0.645)	0.10 (ns: p=0.540)
	Summary	<i>Both Mix A and Mix B increase GHA levels</i>	<i>Both Mix A and Mix B increase GHA levels</i>	<i>GHA levels not affected by either Mix A or B</i>	<i>GHA levels not affected by either Mix A or B</i>
Genotype	allele present vs allele absent	0.51 (*p=0.021)	0.38 (ns: p=0.071)	-0.17 (ns: p=0.458)	-0.24 (ns: p=0.222)
	Summary	<i>GHA levels are higher with allele present</i>	<i>GHA levels and allele presence/absence not related</i>	<i>GHA levels and allele presence/absence not related</i>	<i>GHA levels and allele presence/absence not related</i>
Challenge (Mix A or Mix B) x Genotype (allele present or absent):	Mix A vs Placebo	-0.53 (*p=0.041)	-0.46 (ns: p=0.061)	-0.23 (ns: p=382)	0.01 (ns: p=0.959)
	Mix B vs Placebo	-0.40 (ns: p=0.134)	-0.23 (ns: p=0.338)	0.12 (ns: p=0.662)	0.20 (ns: p=0.389)
	Summary	<i>Absence of allele enhances GHA response to Mix A but not to Mix B</i>	<i>GHA response to Mix A and B not affected by allele</i>	<i>GHA response to Mix A and B not affected by allele</i>	<i>GHA response to Mix A and B not affected by allele</i>

TABLE 7. Mixed Models analyses of interactions between Genotype and Behavioral Responses (GHA) to Additive Mix A and Mix B versus Placebo in the ≥85% consumption sub-sample of 8/9 YO Children with All Potential Confounds Controlled

		[Effect Size (p value)]			
		HNMT Thr105Ile	HNMT T939C	COMT Val108Met	ADRA2A C1291G
Additive Challenge	Mix A vs Placebo	0.11 (*p=0.046)	0.19 (*p=0.003)	0.08 (ns: p=0.379)	0.11 (ns: p=0.104)
	Mix B vs Placebo	0.19 (*p=0.001)	0.25 (*p<0.001)	0.14 (ns: p=0.151)	0.14 (*p=0.036)
	Summary	<i>Both Mix A and Mix B increase GHA levels</i>	<i>Both Mix A and Mix B increase GHA levels</i>	<i>GHA levels not affected by either Mix A or B</i>	<i>Only Mix B increases GHA levels</i>
Genotype	allele present vs allele absent	0.01 (ns: p=0.956)	0.18 (ns: p=0.089)	0.12 (ns: p=0.295)	0.05 (ns: p=0.649)
	Summary	<i>GHA levels and allele presence/absence not related</i>	<i>GHA levels and allele presence/absence not related</i>	<i>GHA levels and allele presence/absence not related</i>	<i>GHA levels and allele presence/absence not related</i>
Challenge (Mix A or Mix B) x Genotype (allele present or absent):	Mix A vs Placebo	-0.10 (ns: p=0.403)	-0.24 (*p=0.021)	0.02 (ns: p=0.874)	-0.05 (ns: p=0.607)
	Mix B vs Placebo	-0.24 (*p=0.050)	-0.23 (*p=0.026)	0.02 (ns: p=0.865)	-0.004 (ns: p=0.967)
	Summary	<i>Absence of allele enhances GHA response to Mix B but not to Mix A</i>	<i>Absence of allele enhances GHA response to both Mix A and Mix B</i>	<i>GHA response to Mix A and B not affected by allele</i>	<i>GHA response to Mix A and B not affected by allele</i>

There were no significant main effects of any genotype on GHA at baseline (T0) in either age group of children (data not shown). During the challenge study, only one genotype, HNMT Thr105Ile, in the 3/4YO children showed a significant (p=0.02) main effect on GHA levels, which were higher with Thr105Ile/*present* compared to Thr105Ile/*absent*. The relevance of this is not known.

The results of the interaction analyses showed that the COMT Val108Met and ADRA2A c1291g polymorphisms apparently had no modulating influence on the effects of AFCAs (Mix A and B) on GHA levels in either the 3/4YO or 8/9YO children. However, significant modulating effects were found for both HNMT Thr105Ile and the HNMT T939C polymorphisms in both the 3/4YO and 8/9YO children.

Specifically, for the 3/4YO children a modulating effect of the HNMT Thr105Ile/*present* genotype was found which significantly (p=0.04) reduced the adverse effects of Mix A, that is, fewer *hyperactivity profile behaviors* were seen in response to Mix A challenge [see *Reviewer Comment (b)*, below]. A similar, but nonsignificant (p=0.06), moderating effect of the HNMT T939C/*present* genotype on

the adverse effects of Mix A was noted. Neither the HNMT Thr105Ile nor the HNMT T939C genotype significantly influenced the effects of Mix B in the 3/4YO children.

For the 8/9YO children the HNMT Thr105Ile/*present* genotype significantly ($p=0.05$) reduced the adverse behavioral effects of Mix B, but not Mix A. The HNMT T939C/*present* genotype had an even greater modulating influence, significantly reducing the adverse effects of both Mix A ($p=0.02$) and Mix B ($p=0.026$).

In the discussion of these genotype findings the principle investigators suggest a link between histamine and hyperactivity with certain polymorphisms in the HNMT gene moderating behavioral responses to the mixture of food colorings and the benzoate preservative present in Mix A in the 3/4YO children and in Mix A and B in the 8/9YO children. They also suggest that the current focus on dopamine in studies of ADHD needs to be extended to histamine. HNMT polymorphisms impair histamine clearance. The presence of H3 receptors in the brain provides a possible mechanistic explanation for the interactive effects found. Many environmental factors can increase histamine release, including infections as well as many food items and certain artificial food colors. The authors indicate that this would explain the frequent claim that food allergy/intolerance is a cause of hyperactivity and the effects of infections in aggravating aberrant behavior. This clearly indicates a potential target for therapeutic intervention in ADHD focused on the H3 receptor [see *Reviewer Comment (c)*, below].

6. Reviewer Comments: Modulation of AFCA behavioral effects by genetic polymorphisms – Study Findings

(a) Variability in finding of significance for main effect of challenge (Mix A and Mix B) on behavioral responses (GHA) with repeated Mixed Models analyses for different genotypes raises questions about replicability of the Mixed Models analyses and uncertainties about the findings of significant treatment effects in the primary study.

For each age group, the $\geq 85\%$ consumption sub-sample of subjects identified in the primary study was used for determining the modulating influence of four different genotypes on the behavioral effects of treatment (Mix A and Mix B). Four separate datasets each comprised of virtually all of the same subjects were used for the four genotype analyses. Each of the four datasets used for the four genotype analyses was initially tested (Mixed Models analysis) for the main effect of challenge on GHA levels, independent of whether the genotype allele was present or absent, and subsequently tested for Challenge x Genotype (polymorphism *present* or *absent*) interactions. Effectively, the same Mixed Models analysis for a main challenge effect was repeated four times using separate datasets comprised of basically the same population of experimentally treated subjects; thus, comparable statistical findings with regard to main treatment effects (independent of genotype being present or absent) should occur across all four genotype analyses. In fact, however, the statistical findings regarding main challenge effects were notably inconsistent across the four genotype analyses. In the 3/4YO children significant challenge effects were found for Mix A and Mix B in the analyses of the two HNMT Thr105Ile and HNMT T939C datasets, but not in the analyses (using essentially the same sample of children) of the COMT Val108Met and ADRA2A C1291G datasets. Comparably, in

the 8/9YO children significant challenge effects were found for Mix A and Mix B in the analyses of the HNMT Thr105Ile and HNMT T939C datasets and for Mix B in the analyses of the ADRA2A C1291G dataset, but not in the analyses of the COMT Val108Met dataset. Even though these analyses were conducted using only the $\geq 85\%$ consumption sub-sample of children, such variability in the statistical findings raises questions regarding the replicability of the Mixed Models method at least as used in the determination of the modulating influence of genotype on the behavioral effects of AFCA. A specific evaluation by a statistician of the adequacy and replicability of the statistical procedures used would be appropriate.

(b) The moderating influence specifically of the Thr105Ile/present genotype decreasing the behavioral effects of Mix A in 3/4 YO children appears contradictory to the main effect of genotype on GHA levels.

The association of Thr105Ile/present with a reduction in the behavioral effects of Mix A in 3/4 YO children appears contradictory, since a main effect analysis showed this same genotype, Thr105Ile/present, to be significantly associated with higher overall GHA levels compared with Thr105Ile/absent genotype. The investigators attempted to address this by suggesting that the two histamine risk alleles in this study have two actions. The first is to influence the overall level of GHA, significantly for the younger children, and second to make the children more vulnerable to the effects of AFCAs on behavior. They add that the role of genes in influencing behavior needs to be understood not by just their main effects of raising levels, for example, of hyperactivity (*hyperactivity profile behaviors*) but also by the interplay both with each other in gene-gene interactions and also by interactions with environmental factors such as diet. The above suggested effects for the histamine risk alleles are not completely consistent with the finding that there were no main effects for either histamine polymorphism on GHA levels in the 8/9 YO children, yet both Thr105Ile and T939C were shown to have had significant modulating effects on the behavioral responses to AFCA challenges.

(c) The findings from the analysis of the genotype component study data which involved children from the general population were inappropriately extrapolated to ADHD children.

The suggested utility of genotype results from this study of behavior in a general population of children to studies dealing with the specific condition of ADHD is an inappropriate extrapolation. The present study attempts to provide some information related to whether any effects of AFCAs on behavior are modulated by genetic differences between children in a normal population. This is not the same as a population of ADHD children. ADHD is a specific neurologic disorder characterized by dysfunctional behavior and is not the extreme end of a biological continuum of normal behavior.

7. Additional Secondary Questions – Study Findings

(a) Consistency in response between Mix A and Mix B.

The challenge mixes differed from one another both in terms of the artificial food colors included and in the doses. It is therefore difficult to interpret differential responses by individual children to the two mixes. The distribution of the GHA scores for both mixes at both ages was normal, the effects were on a continuum and there was no immediate evidence in the distribution pattern of a sub-group of children who were distinctively responsive to the mixtures.

(b) Dose-response relationship.

This study was not designed to investigate the effect of dose on response. In fact the variations in dose taken for each mix were not under experimental control but rather arose both from different levels of color additives in each mix and from differential compliance with the consumption of the stipulated dose in the challenge drink. This means that the interpretation of the dose-response relationship is problematic not least in terms of ambiguity in the direction of the effects.

(c) Difference between the 3/4 year old and 8/9 year old children in changes in GHA over study period.

Following the design requirements from the Food Standards Agency, after a baseline period on normal diet, the children were placed on a withdrawal diet without color additives or benzoate preservatives and simultaneously started on a placebo drink. This was done to minimize the placebo effects such that throughout the study the children were receiving a drink of some kind. This meant that the effects of withdrawal were confounded with those of placebo. The pattern of changes in GHA over the period of the study for the 3/4YO children was such that the *hyperactivity profile behaviors* remained below baseline (TO) levels (normal behavior for general population children on regular diets) even though they were exposed to the active challenge mixes for 2 of the 6 weeks of the study period. This was interpreted as the effects of withdrawal which reduced the *hyperactive profile behaviors* such that subsequent challenges to raise levels of these behaviors were counteracted. For the 8/9YO children the GHA tended to increase above baseline level even during periods of placebo challenge. This was deemed due to the inclusion of the Continuous Performance Test (CPT) for the 8/9YO children. This component of the GHA, which was absent for the 3/4YO children, showed progressively worsening scores over the study period, resulting in progressively increasing GHA scores. All other component measures of the 8/9YO children's GHA remained below or close to the baseline values. The principle investigators noted that the CPT time effects were controlled in the Mixed Models analyses reported for this study [see *Reviewer Comment in Section IV,B.1, above*].

V. Study Investigators' Overall Conclusions of Study

The investigators asserted that this study provided evidence that adverse effects of certain mixtures of artificial food colors and benzoate preservative (AFCA) on hyperactivity can be identified in general population samples of 3/4 year old and 8/9 year old children under certain circumstances. They also contended that these findings replicate the adverse behavioral effects of Mix A previously reported on a large sample of 3 year old children from the general population. In their opinion this evidence collectively provides support for the case that certain food additives may exacerbate hyperactive behaviors (*inattention, impulsivity and overactivity*) in some groups of children. The size of the effects of the AFCA on the average hyperactivity score is lower than that reported for clinical samples and the level of individual variation in response was high. The investigators assert that there are major genetic influences on hyperactivity and this study has shown that differential sensitivity to AFCA resulting from selected genetic polymorphisms is one means by which genetic influences on hyperactivity may be mediated. The investigators consider that these findings demonstrate that adverse effects are not found only in those children at the extreme of hyperactivity, namely those diagnosed with ADHD, but can also be found in the general population and across a range of severity of hyperactivity.

VI. Reviewer Conclusions

The primary hypothesis tested in the present study was that mixtures of certain artificial food colors and sodium benzoate increase the mean level of *hyperactivity profile behaviors* of inattention, impulsivity and overactivity in children from the general population. It should be noted that the nature of the “hyperactivity” behaviors and the magnitude of change reported in this study are not associated with or indicative of ADHD, but rather refer to behaviors occurring within the general population. No information was provided in the study to suggest that the behavioral changes noted were adverse, detrimental or maladaptive in any demonstrable manner. Two age groups from the general population were used, a 3/4 YO group and an 8/9 YO group. The study design was a double-blind within-subject cross-over challenge with two active challenge items and a placebo, each administered in randomized order in a fruit juice drink daily for a one week period. The active challenges were two mixtures (Mix A and B), each containing artificial food colors and sodium benzoate. Mix A and B differed from each other in the quantity and composition of the color additives used, but the level of sodium benzoate remained constant. The active mixtures used to challenge the 8/9 YO children, relative to the 3/4 YO children, had appropriately higher quantities of food colors (but not higher levels of sodium benzoate) to account for their larger size. The placebo was a mixture of unspecified fruit juices. A range of weekly behavior measures was used to detect treatment related effects, including parent ratings, teacher ratings, classroom observation ratings, and, for the 8/9 YO children only, a continuous performance test (CPT). However, the primary outcome measure used in this study to assess treatment related changes in *hyperactivity profile behaviors* was a “global aggregated measure of hyperactivity” (GHA) which was derived by combining the standardized z-scores from the different sources of behavior measures. A high GHA indicated greater *hyperactivity profile behaviors*.

Based on the analysis of the *whole sample* of children (considered the primary analysis), the findings from the primary study were that challenge with Mix A of artificial food colors and benzoate preservatives elicited statistically significant increases in GHA levels of 3/4 YO children and that challenge with Mix B elicited statistically significant increases in GHA levels of 8/9 YO children. Additional analyses were conducted to assess the effects of both challenge Mixes on each individual (disaggregated) component behavior measure. The principle investigators attempted to minimize the significance and utility of the individual (disaggregated) behavior measures as being less sensitive and less reliable than the global aggregate measure, GHA. This reviewer does not agree, but is of the opinion that, in order to fully interpret the significance of the treatment-related findings in the primary study, it is important to consider how each of the component behavior measures contributed to the significant treatment effects on the overall GHA levels. Analyses of the disaggregated behavior measures were sufficiently sensitive to identify a number of significant behavioral changes. The finding of complementary significant treatment related changes across several of the various behavior measures would not only provide confirmation of biologically relevant changes but would also provide specific information about the treatment effects that could enhance interpretation of the study findings. The analyses of the individual (disaggregated) component behavior measures for the *whole sample* of subjects indicated that the parental rating was the major contributor for the primary effects of Mix A on GHA levels in 3/4 YO children, although the parental rating measure alone did not show a statistically significant effect of Mix A challenge. In the *whole sample* of 8/9 YO children the parental rating and CPT scores were the major contributors for the primary effects of Mix B on GHA levels but only the parental rating and not the CPT scores showed a statistically

significant treatment effect for Mix B. In neither age group of children were there any notable changes in teacher or classroom observation ratings, indicating that these behavior measures did not confirm the overall treatment effects on GHA levels based on parental ratings. A previous study by Bateman et al (2004), using a large sample of 3 year old children also from the general population, found significant behavioral effects of Mix A based specifically on parental ratings but, similar to the present study, these effects were not confirmed by a clinical behavioral evaluation. Contrary to the contention of the principle investigators, the previous study's significant findings were not completely replicated in the present study, since, as noted above, the analysis of the individual disaggregated component behavior measures revealed that the parental ratings (the primary contributor to the significant effect of Mix A on GHA levels) showed only a non-significant trend for behavioral changes with Mix A challenge.

A rationale was suggested by the principle investigators to explain why treatment effects in the primary study based primarily on the parental ratings were not confirmed by teacher or classroom observation ratings. The principle investigators interpreted trends (findings were not statistically significant) from an "acute challenge" study using 8/9 YO children to suggest a rather short onset time (less than an hour) for the appearance of treatment effects from Mix B challenge. The principle investigators concluded that, since all of the children in the primary study consumed their challenge drinks at home, the short onset time for treatment effects to appear would suggest that it was more likely for the parents in the home setting, rather than teachers or classroom observers, to have seen the behavioral changes in the children. In the opinion of the reviewer, this scenario does not adequately resolve the issue of parental ratings not being confirmed by teacher or classroom observations for two basic reasons. First, the investigators did not consider the importance of when the children consumed their challenge drinks. If the drinks were consumed before going to school the teachers/classroom observer might have been more likely than parents to see the treatment effects. The opposite would be more likely if drinks were consumed after school. Unfortunately, the study report stated only that the 8/9 YO children usually consumed their drinks after returning from school and the 3/4 YO children consumed their drinks either prior to or upon returning from their school setting, but no specific information was given. Second, while the "acute challenge" study suggested a short onset time for the appearance of treatment effects, that study was not designed to determine how long those treatment effects would last. It is conceivable that treatment effects, if any occurred, could have persisted into the day after challenge, when the children returned to the school or early year setting, and if so, should have been detected by the teachers and classroom observers. None of the findings from the "acute challenge" study, therefore, help explain why the overall effects of challenge on GHA levels appeared to be based primarily on the parental ratings for the *whole sample* of both aged groups of children, along with the CPT scores for the 8/9YO children, with only nominal, if any, treatment effects being detected with teacher ratings or classroom observation ratings. Numerous alternative explanations for the lack of confirmation of parental detection of treatment effects by other behavior measures in either age group of children could be speculated. For example, the specific or subtle behavioral changes the parents detected were either not detectable with the other behavior measures or were not expressed in the classroom or other test environments. Consideration should, however, be given to possible alternative explanations, for example that the blind may somehow have been broken, or that the parental findings were simply statistical false positives. In the absence of an adequate explanation, the fact that there was no confirmation of trends for treatment effects between parental ratings and the other behavior measures lessens confidence in the relevance and reliability of the parental based findings.

There are several data analysis issues that may also affect the confidence in the finding of statistically significant overall effects in the primary study including significant effects of Mix A on GHA levels in the 3/4 YO children, and of Mix B in the 8/9 YO children. One issue involves the fact that, in the statistical analyses for treatment related changes in overall GHAs and in the component (disaggregated) behavior measures for both age groups, the finding of statistically significant effects varied notably depending upon whether the *whole sample* or either of the sub-samples (*>85% consumption* group or *complete case* group) was used for the analysis. As noted by the principle investigators, variability in the statistical outcome between the whole sample and sub-sample analyses may indicate that non-compliance in consuming the challenge drinks or the method used to handle missing data may have affected the pattern of results. But, in describing the results the investigators did not discuss or clarify the impact or relevance of these problematic aspects of the data analyses on the confidence in the statistical significance of the study findings and their interpretation. This uncertainty lessens confidence in the reliability and relevance of the statistically significant study findings. A second issue concerns the inconsistency in finding significant main effects of challenge (Mix A and Mix B) on overall behavioral responses (GHA levels) in the analyses of the genotype data (a secondary research component of this study discussed below). In each age group of children a Mixed Models analysis was conducted to test for the influence of each of four genotypes on the behavioral effects of AFCAs. Since the analysis of each of the four genotypes included a test for main effect of challenge (Mix A and Mix B) on GHA levels, the analysis for a main challenge effect was repeated four times using basically the same group of subjects (only the $\geq 85\%$ *consumption* sub-sample was tested). While significant main challenge effects were occasionally detected, these effects were not consistently detected across the four repeated analyses. Although these particular repeated analyses were conducted using the same $\geq 85\%$ *consumption* sub-sample of children, the inconsistency in finding significant main challenge effects raises questions regarding the replicability of the Mixed Models statistical analysis. Since the Mixed Models analysis was used to analyze the primary study data, there are also uncertainties about the replicability and the relevance of the statistically significant challenge effects in the primary study. A specific evaluation by a statistician of the adequacy and replicability of the statistical procedures used for the primary study and the genotype analyses would be appropriate.

The present study also addressed a secondary question involving the possibility that the behavioral effects of AFCAs (Mix A and Mix B) may be moderated by genetic differences among children in a normal population. Four selected genetic polymorphisms were analyzed using cheek cell DNA from all children of both age groups. The genetic data from the $>85\%$ *consumption* sub-sample of children were evaluated for interactive effects on the AFCA related changes in behavior. Suggestive findings were presented that supported some level of genetic influence (specifically two histamine N-methyltransferase gene polymorphisms, Thr105Ile and T939C) on the sensitivity to the behavioral effects of certain AFCAs in children from the general population. Due to the fact that no methodological details were provided for the genotyping analyses and the questions regarding the statistical analysis used for these data, it was difficult to assess the adequacy of the procedures used in this analysis or the reliability of the findings. These interesting but preliminary genotype findings should be replicated and extended in a more focused and well-designed study with more detailed information about the procedures used.

Overall, the primary study findings are suggestive of low level behavioral effects of AFCAs on behavior in 3/4 YO children (Mix A) and 8/9 YO children (Mix B), limited to detection based primarily on parental ratings and possibly CPT scores for the older children. However, due to the absence of confirmation of treatment effects between parental ratings and other sensitive

behavior measures together with the concerns about the data analyses described above and various procedural weaknesses (outlined in the section on Study Weaknesses, below), it is the opinion of this reviewer that there is questionable confidence in the reliability and biological relevance of the primary findings from this study.

VII. REVIEWER NOTED STUDY STRENGTHS

- (1) The recruitment procedure enabled the selection of two study groups of children, one comprised of 3/4 YO and the other 8/9 YO, drawn from a general (normal) population of children.
- (2) Schools (8/9 YO children) and Early Year Settings (3/4 YO children) were selected to ensure that the study samples reflected the full range of socio-economic backgrounds of age-matched children in the study area.
- (3) To further assess how representative the sample was in terms of behavior, the teachers in the participating Early Year Settings and schools completed a behavioral profile questionnaire for all 3/4 YO and 8/9 YO children. The study sample was found to be representative of the general population of age and gender-matched children in terms of the teacher behavioral ratings.
- (4) The selection procedure was structured such that for each age group there were no significant differences in socioeconomic or other background characteristics between any of the three sample groups used for statistical analyses or between groups of children assigned to receive the challenge drinks in different orders over the 6 week study period.
- (5) The experimental study was appropriately designed as a within-subject crossover with three challenge treatments, two additive mixes (Mix A and Mix B) and a placebo administered in fruit juice under double-blind conditions.
- (6) Multiple behavioral measures were used to detect any treatment effects. The measures included standardized parent ratings, teacher ratings, classroom /playroom observations, and a continuous performance task (8/9 YO children only).
- (7) Several correlative secondary research questions were also addressed in this study, including: 1) whether challenge related behavioral effects are seen across the different sources of behavior measures: parent ratings, teacher ratings, direct classroom/playroom observation of behavior, and continuous performance testing (8/9YO only); 2) whether behavioral effects of AFCAs (Mix A and Mix B) may be modulated by genetic differences between children in a normal population; and 3) whether it is possible to demonstrate short term changes in behavior immediately after single dose acute challenge with a mix of AFCAs and to identify metabolic factors that may mediate such responses.

VIII. REVIEWER NOTED STUDY WEAKNESSES

There are a number of weaknesses that impact confidence in the study findings to various degrees. Specific weaknesses or shortcomings of this study are presented below in relation to (1) the procedures (Materials and Methods) used in the conduct of this study and (2) the analyses of the

study data and interpretation of the study findings.

A. Procedural Weaknesses

- (1) There is questionable accuracy of the teacher and parent pre-study behavioral ratings. The teachers and parents completed behavioral questionnaires prior to the start of the study to indicate frequency of certain behaviors for individual children over the previous 6 months. The accuracy of ratings based on a recall of behaviors over a period of 6 months is highly questionable, particularly for teachers rating each of multiple children in a classroom or playroom setting.
- (2) The two active challenges (A and B) used in this study were both mixtures of four color additives plus sodium benzoate. The use of chemical mixtures as challenge materials precludes identifying which specific compound(s) within the mixtures might be responsible for any treatment-related effects. In order to identify the specific active compound(s), subsequent studies would be needed in which each chemical is individually is used as a challenge substance, preferably at several dose levels.
- (3) No explanations were given for the use of different color additives between Mix A and Mix B making the composition of the two mixes different; for the disproportionate increases in the levels of the color additives in the mixes (A and B) used for the 8/9YO children relative to the mixes (A and B) used for the 3/4YO children; and for the absence of adjusting the levels of sodium benzoate between age groups of children effectively lowering the dose level of sodium benzoate for the older children. Each of these procedural elements complicates the comparative assessment of treatment effects between Mix A and Mix B within or between age groups of children and limits the usefulness of such comparisons.
- (4) The composition of the placebo was not clearly described. The placebo is a critical component of this, or any, challenge study and it would be important to have information about the types of fruit juice used (commercial or fresh). If commercial, what brands were used and did they contain any types of additives? If fresh, how was the juice made and when was it prepared relative to use in the study? Such information about the placebo source and composition are very important in helping to assess the outcome from this study.
- (5) The parents were not asked whether they could tell the difference between the active and placebo drinks, as a check that the blinding was intact in the home environment.
- (6) The detection of treatment related behavioral effects in the primary study was based principally on parental ratings, but the specific behavioral elements in the parent questionnaire that were most affected were not identified.
- (7) The investigators did not provide comparative standard teacher/parent behavioral ratings or COC ratings for ADHD and non-ADHD children, with which to help gauge the significance of the ratings for the children in this study or to help identify responders versus non-responders.
- (8) Each child's classroom observation score for an entire week was based on only three 8-minute observation periods (total of 24 minutes of observation per week). It is

questionable whether observing a child for only 8 minutes three times a week is an adequate sampling frequency or observation duration to provide a reliable measure of representative classroom behavior over an entire week.

- (9) In the “acute challenge phase” of this study the disaggregated behavior measures comprising the Hyperactivity Index (HI) were not analyzed separately. The component behavior measures, CPT and independent behavior ratings, were aggregated to form the Hyperactivity Index (HI). However, these component behavior measures were not analyzed separately (disaggregated) to help determine whether changes in CPT performance or the behavior rating contributed more, less or equivalently to any treatment related changes in the overall aggregate HI scores.
- (10) The Mixed Models analyses of the “acute challenge” data did not appear to control for potential confounding factors. In the absence of controlling for confounding variables there is low confidence in the reliability of the statistical analyses of the “acute challenge” study findings.
- (11) No methodological details were provided for the genotyping analyses. The investigators did not provide any description of the materials used or any procedural details about the manner in which DNA samples were taken, at what period during the main study the samples were taken, how the samples were maintained, or how the samples were analyzed for SNPs. Without such methodological information it is difficult to assess the adequacy of the procedures used in this analysis or the reliability of the data collected.
- (12) With reference to the secondary question of whether genotype may have a modulating influence on the behavioral effects of AFCA treatment, one important aspect that was not considered is the comparative genotype in the sub-set of AFCA “responders” versus “non-responders”. Determining the relative distribution of the presence or absence of each of the polymorphic alleles for “responders” and “non-responders” would have complemented the determination of whether the presence or absence of certain gene polymorphisms correlate with the adverse effects of AFCA on behavior.

B. Weaknesses Regarding Data Analyses and Interpretation of Study Findings

- (1) In the statistical analyses for treatment related changes in overall GHAs or in the component (disaggregated) behavior measures for both age groups of children, the finding of statistically significant effects varied notably depending upon whether the *whole sample*, the $\geq 85\%$ *consumption* sub-sample, or the *complete case* sub-sample of children was used for the analysis. For example, in the analysis of overall GHAs when either the *whole sample* or the $\geq 85\%$ *consumption* sub-sample of 8/9YO children was used for analysis, only Mix B had a significant effect. However, when the *complete case* sub-sample of 8/9YO children was used, both Mix A and Mix B had significant effects on GHA scores. In the analyses of the component (disaggregated) behavior measures, when the *whole sample* of children was used, there were no significant effects for the 3/4YO children and for the 8/9YO children only the parent ratings were significantly affected with Mix B challenge. None of the teacher ratings or classroom observation ratings (both age groups) or CPT scores (8/9YO) were significantly

affected. However, when the analyses used the *complete case* sub-sample of children, parental ratings were significantly affected only for the 3/4YO children with Mix A, and CPT scores for the 8/9 YO children were significantly affected only with Mix B. The teacher and classroom observation ratings were still not significantly affected. This variability in statistically significant outcomes may indicate that non-compliance in consuming the challenge drinks or the method used to handle missing data in the analyses had affected the pattern of results. But in describing the results the investigators did not discuss these problematic aspects of the data analyses, in particular what relevance they have to the interpretation of the study findings. This lessens confidence in the reliability of the statistically significant study findings.

- (2) Regarding the decreasing CPT performance of the 8/9 YO children over the study period, little detailed information was presented, but the study investigators concluded that this decline was due to the children becoming bored with the repeated conduct of this inherently tedious task. Sufficient additional information should have been provided to support this conclusion.
- (3) None of the confounding variables, which were controlled in the data analyses of the primary study, were incorporated into the Mixed Model methods used for analysis of the “acute challenge” data. Since potential confounds were not controlled in the data analysis, this calls into question the reliability of the findings from the “acute challenge” component study.
- (4) The inconsistent finding of significant main effects of challenge (Mix A and Mix B) on behavioral responses (GHA) in the analyses of the genotype data raises questions about replicability of the Mixed Models analyses and uncertainties about the findings of significant treatment effects in the primary study. In each age group of children (3/4YO and 8/9YO), to test for the influence of four genotypes on the behavioral effects of AFCAs, a Mixed Models analysis was conducted on the data of four sets of subjects, taken only from the $\geq 85\%$ consumption sub-sample, with each set comprised of virtually the same subjects. Since the analysis of each of the four genotypes included a test for main effect of challenge (Mix A and Mix B) on GHA, the analysis for main challenge effects was repeated four times using basically the same group of subjects. If a statistically significant main challenge effect of treatment was present, it would have been expected to be found for each of the four sets of subjects. While a significant main challenge effect was present, this effect was not consistently detected across the four sets of subjects. In the 3/4YO children main challenge effects for both Mix A and B were found in only two of the four analyses; in the 8/9YO children significant main effects were found for Mix A in two of the four analyses and for Mix B in three of the four analyses. This inconsistency in finding significant main challenge effects raises questions regarding the replicability of the Mixed Models statistical analysis. Since the Mixed Models analysis was used to analyze the primary study data, there are uncertainties about the replicability and the relevance of the statistically significant challenge effects in the primary study. A specific evaluation by a statistician of the adequacy and replicability of the statistical procedures used for the primary study and the genotype analyses would be appropriate.
- (5) The rationale offered by the investigators to explain why treatment effects in the primary study appeared to be based primarily on the parental ratings with little

contribution from the other behavior measures was based on an inaccurate interpretation of the acute study findings. The principle investigators contend that the findings from the acute challenge study suggest that the 8/9 YO children who respond to the ingestion of Mix B challenge do so within a short period of time, an hour. Since the children in the primary study consumed their challenge drinks at home, the findings from the acute challenge study would suggest that it was more likely for the parents in the home setting, rather than teachers or observers, to have observed the behavioral changes and may be an explanation as to why the effects were detected principally by the parental ratings rather than the school or early years settings measures. The trends in the acute challenge study (although not statistically significant) may suggest that the onset of response following ingestion of Mix B challenge occurred within a short period of time, an hour. However, the acute challenge study is not designed to determine the duration of any treatment effects. Consequently, though some of the children may have consumed their challenge drinks at home after returning from school (or from the Early Year Setting) and experienced the onset of treatment effects shortly thereafter, the acute study provides no information about how long those effects may have lasted. It is possible that treatment effects, if any occurred, could have persisted at least into the day after challenge, when the children returned to the school or the Early Year Setting environments.

- (6) The use of the three sample groups of children for analysis of the component (disaggregated) behavior measures resulted in variable statistical outcomes but relevance of this to data interpretation was not discussed in the technical report. As was apparent in the analyses of the change in GHAs for the component behavior measures, there is notable variability in the occurrence of statistically significant effects across subgroups. Depending upon whether the *whole sample*, the $\geq 85\%$ consumption sub-sample or the *complete case* sub-sample of children was used for the analysis, the outcome of significant effects on certain component behaviors is notably different. For example, based on analysis of the component behaviors using the *whole sample* of children, there were no significant effects for the 3/4YO children, and for the 8/9YO children only the parent ratings were significantly affected with Mix B challenge. None of the teacher ratings or classroom observation ratings for either age groups or CPT scores (8/9YO) were significantly affected. However, when the analyses used the *complete case* sub-sample of children, parent ratings were significantly affected only for the 3/4YO children with Mix A, and CPT scores were significantly affected but only for the 8/9YO children with Mix B. The teacher and classroom observation ratings were still not significantly affected. Although this obvious variability in significant treatment effects across subgroups of subjects may indicate that non-compliance in consumption of challenge drinks or the method of handling missing data affected the pattern of results, the investigators do not discuss the impact of this problematic aspect of the data analyses on interpretation of the study findings.
- (7) There was no basis presented for considering the behavioral changes in this study as necessarily “adverse” or “deleterious”. In the previous study by Bateman et al (2004) treatment effects for Mix A were found based on the daily parental ratings, but were not confirmed in the weekly clinical assessments by research psychologists using validated tests. In the present study the *whole sample* analyses showed significant effects of Mix A on GHA levels for the 3/4YO children and significant effects of Mix B on GHA for the 8/9YO children. Analysis of the disaggregated component behavior measures for

the *whole sample* of subjects (see Results section) indicated that parental ratings for both age groups of children and CPT scores for the 8/9YO children (not measured in the younger children) were the major contributors to the significant AFCA effects on the overall GHA levels. Notably, there appeared to be little suggestion of treatment effects based on either the teacher ratings or the classroom observation ratings, indicating that Mix A and Mix B did not elicit behavioral changes that adversely affected the early years setting (3/4YO children), school (8/9YO children) or observation classroom (both age groups) environments. There was no information provided in either study to suggest that the changes in behavior based on parental reports were adverse, detrimental or maladaptive. The children's behavior under challenge conditions appeared to be within the range of behavioral levels exhibited by the general population of age-matched children.

- (8) The suggested utility of genotype results from this study of behavior in a general population of children to studies dealing with the specific condition of ADHD is an inappropriate extrapolation. The present study attempts to provide some information related to whether any effects of AFCAs on behavior (*'hyperactivity profile behaviors'* of over activity, inattention and impulsivity) are moderated by genetic differences between children in a normal population. This is not the same as a population of ADHD children. ADHD is a specific neurological disorder characterized by dysfunctional behavior and is not the extreme end of a biological continuum with normal behavior.

IX. Applicability to Assess Risk or to Support Regulatory Action

The primary findings in the present study showed that 1 week of daily challenges with certain mixtures of artificial food colors and sodium benzoate may produce low level behavior effects in 3/4 YO and 8/9 YO children from the general population, limited to detection based primarily on parental ratings and possibly continuous performance test scores. There was no indication that any of these behavioral findings constituted clear adverse effects. However, due to the absence of confirmation of the parental rating effects by teacher ratings and classroom observation ratings, along with concerns about data analyses and various procedural weaknesses, there is questionable confidence in the reliability and biological relevance of the primary findings from this study. One particular procedural weakness relevant to regulatory application was the use of chemical mixtures as challenge materials which precludes identifying which specific compound(s) within the mixtures might be responsible for any treatment related effects. Consequently, there would be little, if any utility of these findings to assess risk or to support any specific regulatory decision.

Preliminary genotype findings, which were developed in this study to address a secondary question, tended to support the possibility of some level of genetic influence on the sensitivity to the behavioral effects of certain AFCAs in children from the general population. However, due to data analysis and procedural concerns there are uncertainties regarding the adequacy of the statistical and analytical procedures used and the reliability of the data collected. These uncertainties, together with the preliminary nature of these genotype findings, would indicate little, if any, regulatory utility for these findings.

References

Angello L, Volpe R, DiPerna J, Gureasko-Moore S, Gureasko-Moore D, Nebrig M and Ota K. An assessment of attention-deficit/hyperactivity disorder: An evaluation of six published rating scales. *School Psychology Review* 2003; vol. 32.

Bateman B, Warner JO, Hutchinson E, et al. The effects of a double-blind, placebo controlled, artificial food colourings and benzoate preservative challenge on hyperactivity in a general population sample of preschool children. *Arch Dis Child* 2004; 89: 506-11.

NIH Consensus Conference (Editorial). NIH consensus development conference: defined diets and childhood hyperactivity. *Clinical Paediatrics* 1982; 21: 627-30.

Feingold BF. Hyperkinesis and learning disabilities linked to artificial food flavors and colours. *Am J Nurs* 1975; 75: 797-803.

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

Swanson JM, Sergeant J, Taylor E, Sonuga-Barke E, Jensen PS, Cantwell D. Attention Deficit Hyperactivity Disorder and Hyperkinetic Disorder. *The Lancet* 1998; 351: 429-33.

TECHNICAL DATA EVALUATION REPORT

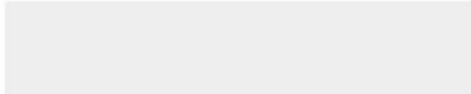
Proposed Association Between
Artificial Food Colors and Attention Deficit Hyperactivity Disorders
(ADHD) and Problem Behaviors in Children

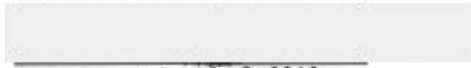
Prepared for
Center for Food Safety and Applied Nutrition
Office of Food Additive Safety
U.S. Food and Drug Administration
College Park, MD 20740-3835

Prepared by
Toxicology and Hazard Assessment Group
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831
Task 2008-30

Primary Reviewer:
Thomas J. Sobotka, Ph.D.

Secondary Reviewers:
Robert H. Ross, M.S., Group Leader

Signature: 
Date: NOV 29 2010

Signature: 
Date: NOV 29 2010

Oak Ridge National Laboratory managed and operated by UT-Battelle, LLC., for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725