


SENDER: COMPLETE THIS SECTION		COMPLETE THIS SECTION ON DELIVERY	
<ul style="list-style-type: none"> Complete Items 1, 2, and 3. Also complete Item 4 if Restricted Delivery is desired. Print your name and address on the reverse so that we can return the card to you. Attach this card to the back of the mailpiece, or on the front if space permits. 		A. Signature X <input type="checkbox"/> Agent <input type="checkbox"/> Addressee	
1. Article Addressed to: <div style="border: 1px solid black; padding: 5px; width: fit-content;"> Marc Hauser, Ph.D. (b) (6), (b) (7)(C) </div>		B. Received by (Printed Name)	C. Date of Delivery
		D. Is delivery address different from item 1? <input type="checkbox"/> Yes If YES, enter delivery address below: <input type="checkbox"/> No	
		3. Service Type <input checked="" type="checkbox"/> Certified Mail <input type="checkbox"/> Express Mail <input type="checkbox"/> Registered <input type="checkbox"/> Return Receipt for Merchandise <input type="checkbox"/> Insured Mail <input type="checkbox"/> C.O.D.	
		4. Restricted Delivery? (Extra Fee) <input type="checkbox"/> Yes	
2. Article Number (Transfer from service label)		7001 2510 0008 6355 4744	
PS Form 3811, August 2001		Domestic Return Receipt 102595-02-M-1540	

CERTIFIED MAIL



7001 2510 0008 6355 4744
7001 2510 0008 6355 4744

U.S. Postal Service
CERTIFIED MAIL RECEIPT
(Domestic Mail Only; No Insurance Coverage Provided)

OFFICIAL USE

Postage	\$	
Certified Fee		
Return Receipt Fee (Endorsement Required)		
Restricted Delivery Fee (Endorsement Required)		
Total Postage & Fees	\$	

Sent To
Marc Hauser, Ph.D.
(b) (6), (b) (7)(C)

Postmark
Here



OFFICE OF
RESEARCH INTEGRITY
3905
2010 JUN 10 A 10 22
HARVARD UNIVERSITY
FACULTY OF ARTS AND SCIENCES

MICHAEL D. SMITH
DEAN

UNIVERSITY HALL
CAMBRIDGE, MASSACHUSETTS 02138

June 9, 2010

CONFIDENTIAL

John E. Dahlberg, Ph.D.
Director, Division of Investigative Oversight
Office of Research Integrity, DHHS
1101 Wootton Parkway, Suite 750
Rockville, MD 20852

Re: (b) (6), (b) (7)(C), Professor Marc Hauser, Harvard University

Dear Dr. Dahlberg,

In accordance with §93.315 of Public Health Service Policies on Research Misconduct at 42 CFR Parts 50 and 93, this letter conveys the Report of the Investigation into research misconduct by Professor Marc Hauser of the Department of Psychology in the Faculty of Arts and Sciences at Harvard University (the "Report").

The enclosed flash drives each contain a full copy of the Report of the Investigating Committee and all attachments, including correspondence by Professor Hauser. Two (identical) copies are enclosed as a precaution in case your office encounters difficulty reading material from one of the drives.

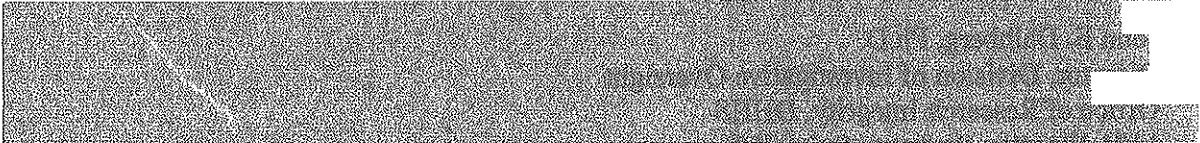
I have reviewed the Report and accept the conclusions of the Investigating Committee that Professor Hauser did commit research misconduct, specifically, falsification of research results, in the eight instances detailed in the Investigating Committee's Report.

The Investigating Committee recognizes that the allegations relating to the first case, referred to as "Theory of Mind," involve events that fall outside the six-year limitation specified in the PHS regulations on research misconduct at §93.105(a). However, I understand that you had indicated, in response to a specific question on this point, that, since the Investigating Committee's Report was to be provided both to the Public Health Service Office for Research Integrity and to the National Science Foundation Office of the Inspector General (because allegations of misconduct involved research supported by both PHS (b) (6), (b) (7)(C)), it was appropriate to provide a single report for both offices; that your office might still have jurisdiction if the questioned research had been reported in proposal submissions or publications

Dr. John E. Dahlberg
June 9, 2010
Page 2

that fell within the six-year window; and that your office would otherwise take into account the six-year limitation in its review of the case.

In light of the Investigating Committee's findings, I have taken actions of appropriate severity and duration regarding Professor Hauser's activities. (b) (6), (b) (7)(C)



In addition, Professor Hauser is required to retract one article (*Cognition* 2002) and to advise the editors of journals that published two other articles (*Science* 2007 and *Proceedings of the Royal Society B* 2007) of the problems with data supporting those two articles' findings, as detailed in the Report, so that the journal editors can determine whether a retraction or correction is required.

Please let me know if you have any questions as you review the Investigating Committee's Report or if I may provide any additional information.

Sincerely yours,



Michael D. Smith

cc: Professor Marc Hauser

OFFICE OF
HARVARD UNIVERSITY
FACULTY OF ARTS AND SCIENCES

Dean R. Gallant **3905**
Assistant Dean JUN 10 A 10:24
for Research Policy and Administration

1414 Massachusetts Avenue, Room 250
Cambridge, Massachusetts 02138
(617) 495-2628 FAX 496-7400

9 June 2010

John E. Dalhberg, Ph.D.
Director, Division of Investigative Oversight
Office of Research Integrity, DHHS
1101 Wootton Parkway, Suite 750
Rockville, MD 20852

Re: (b) (6), (b) (7)(C), Professor Marc Hauser, Harvard University

Dear Dr. Dahlberg,

Since the Report of the Investigating Committee was written, we have learned of a submission by Professor Hauser to the National Science Foundation to support a project entitled "Evolutionary and ontogenetic effects of the acquisition of social preferences in Canids," for the period 1 July 2010 to 30 June 2013. The request was for \$440,432 in support.

The Report of the Investigating Committee states that they were unaware of any pending applications for federal support. I will be grateful if you can amend that information as noted above.

Yours sincerely,



Dean R. Gallant

Dear Sirs,

Following an Inquiry by the Committee on Professional Conduct of Harvard's Faculty of Arts and Sciences, the chair of that Committee, (b) (6), (b) (7)(C), appointed us to constitute an Investigating Committee to assess allegations and evidence for scientific misconduct by Professor Marc D. Hauser (Psychology).

We have considered in detail whether Prof. Hauser committed research misconduct in eight projects. The three committee members have met 18 times, in most cases with Ken Carson, FAS Research Integrity Officer and (b) (6), (b) (7)(C)

In several meetings we were also joined by (b) (6), (b) (7)(C). We have interviewed ten individuals, eight of these in person and two by phone. In addition we have read the transcript of one additional interview conducted by Carson and (b) (6), (b) (7)(C). We have also met with Prof. Hauser and his attorney twice for a total of nine hours. We have read a number of written materials and viewed a number of research related videos from data impounded from his lab and his computers. Several of the people we interviewed also provided written documents for our consideration.

Prof. Hauser provided us with his response to the Committee on Professional Conduct. He has also provided us with a written response to the allegations that our committee has prepared, and he has given us seven letters of support from scientific colleagues, which we read. We also have re-read transcripts of all the interviews that were conducted. Although we read the report of the Committee on Professional Conduct, which ended their work in April 2008, we did not limit ourselves to, or particularly focus upon, the allegations they reviewed. We took as our charge to begin with an open mind.

We have found evidence in support of the allegations. We provide details in this investigative report. The misconduct we have uncovered involves both falsification and reckless handling of data that in our opinion constitute research misconduct as defined by federal regulatory standards.

We present the allegations of research misconduct by Prof. Hauser, along with our assessment of the evidence, in what follows. We conclude this report with a brief discussion of possible exculpatory considerations and interpretations, and a brief assessment of the significance of Prof. Hauser's alleged misconduct within the context of his career as a scientist.

1. Allegations, §93.313 (a)

The Committee has investigated allegations of research misconduct in (b) (6), (b) (7)(C) projects led by Prof. Hauser, and each of the projects will be considered separately below. Almost all of the allegations concern falsification: either miscoding responses of non-human

primates to experimental stimuli, or misrepresenting results in manuscript drafts prepared and submitted for publication, by including incorrect or materially misleading descriptions of experimental results. In several projects, research misconduct is alleged on the basis of evidence that Prof. Hauser knowingly or recklessly destroyed, or failed to maintain, records and that these acts or omissions were a significant departure from the accepted practices of the relevant community.

2a. PHS Support, §93.313(b)

PHS support relevant to the findings detailed in this report includes the following:

P51RR00168-37, a grant to the New England Primate Research Center, supported the initial provision of cotton-top tamarins to the Hauser laboratory, as well as partial support for the animals' upkeep.

CM-5-P40RR003640-13 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), provided funds to maintain the rhesus colony on Cayo Santiago.

NIH/NIDCD award number 5 R01 DC005863 provided support for the tamarin colony and for the work of (b) (6), (b) (7)(C) on the Syl and Seg study.

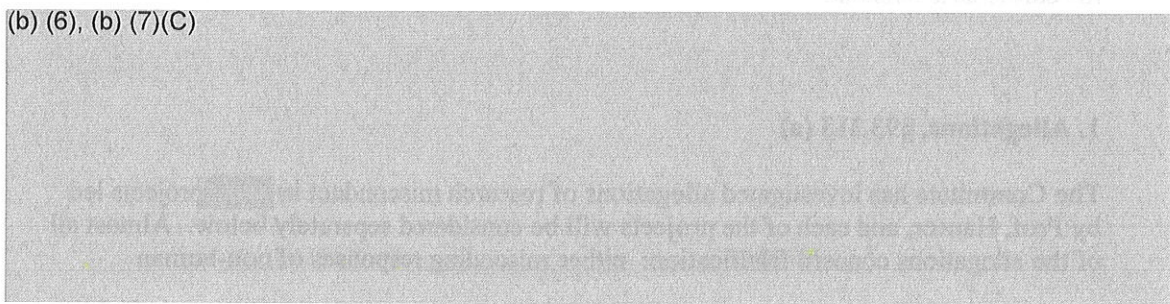
5 F31 MH075298, NRSA fellowship to (b) (6), (b) (7)(C) with Prof. Hauser as sponsor; provided support for research discussed in *Science* and *PRSB* allegations.

The *Cognition* 2002 study cites "a grant to Marcus from the National Institutes of Health," but the Committee understands that Prof. Marcus had no involvement in the conduct of the research or coding of results; his contribution consisted solely of writing parts of the introduction and analysis.

In addition, one funding request to PHS cites research detailed in this report:

"How Nonhuman Primates Perceive Actions: Goals, Motor Experience, and Constraints," NRSA application submitted by (b) (6), (b) (7)(C) (M. Hauser, sponsor) on April 4, 2008 (apparently not funded). The application cites findings from the *Science* and *PRSB* articles and states that (b) (6), (b) (7)(C) "provided Prof. Hauser with several drafts of the grant proposal; the final version is the result of several rounds of revision."

(b) (6), (b) (7)(C)



(b) (6), (b) (7)(C)

3. Institutional Charge, §93.313 (c)

The May 8, 2008 charge letter to the Investigating Committee incorporated the Inquiry Report of the Committee on Professional Conduct dated May 7, 2008 that identified three studies that warranted further consideration: *Cognition* 2002, "Rule Learning by Cotton-Top Tamarins;" *Proc.R. Soc. B* 2007, "Rhesus Monkeys Correctly Read the Goal-Relevant Gestures of a Human Agent;" and *Science* 2007, "The Perception of Rational, Goal-Directed Action in Non-Human Primates" (referred to by the articles reporting results of the studies). The Inquiry Report was sent to ORI on May 7, 2008. The charge letter also advised the Investigating Committee that it "should consider other instances that were presented in the initial allegations, or that may arise in the course of your consideration of the case."

4. Policies and Procedures, §93.313 (d)

The policies under which this investigation was conducted, Harvard University Faculty of Arts and Sciences "Procedures for Responding to Allegations of Misconduct in Research," were provided to ORI with the Inquiry Report, and are also included in the Appendix.

5. Research Records and Evidence §93.313 (e)

- a) Material taken into custody from Prof. Hauser's laboratory and office, summarized on inventory attached. General categories:
 - i) Computer records: forty internal and external hard drives¹, plus other removable media including floppy disks, Zip disks, DAT tapes, etc., containing:
 - (a) Email
 - (b) Data compilations including spreadsheets and statistical program data files; manuscript drafts and associated data files
 - (c) Digitized video and audio
 - ii) Video records
 - (a) Video tape on various tape formats
 - (b) DVDs of video
 - iii) Paper files

¹ All hard drives were copied to new disks and the copies were returned to Prof. Hauser; all original hard drives were retained with other sequestered materials.

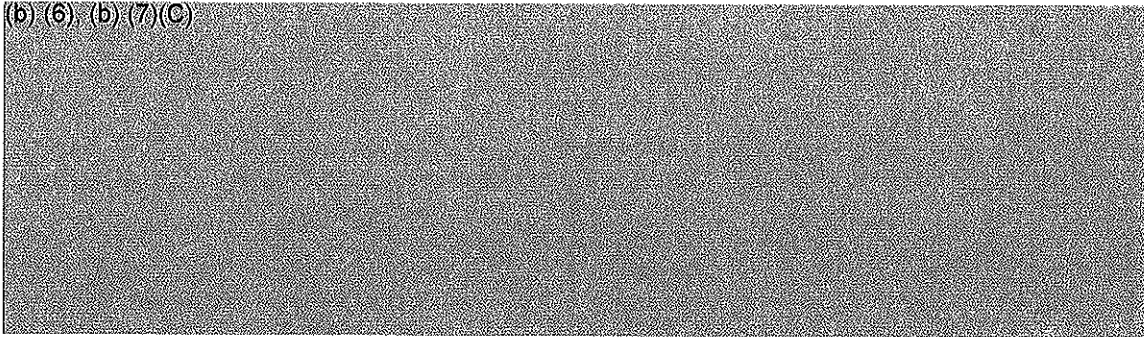
- b) Material provided by others
 - i) DVDs of video clips
 - ii) E-mail
 - iii) Drafts of manuscripts, data for articles
- c) Expert reports: analyses of *Cognition* tape
- d) Material proffered by Prof. Hauser
 - i) Statements from witnesses and experts
 - ii) Data, including video and summaries and analysis, associated with replications or similar studies, and reanalysis of data from studies under question
 - iii) Statements of Prof. Hauser, including written statements and transcriptions of interviews
- e) Interviews of witnesses
 - i) (b) (6), (b) (7)(C)
 - ii)
 - iii)
 - iv)
 - v)
 - vi)
 - vii)
 - viii)
 - ix)
 - x)
 - xi)

Statement of Findings, §93.313 (f)

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)



"Rule learning by cotton-top tamarins"
***Cognition* 86 (2002), p B15-B22**

This paper, published in 2002, and included in Appendix Folder II as item 5, reported that cotton-top tamarins could recognize algebraic "rules" in auditory stimuli, an ability that had earlier been demonstrated only in humans.

The paper reported the results of an experiment conducted from January to March of 2000. A BASF T-180 VHS videotape² contains recordings of the experiment, in which cotton-top tamarins were presented with a series of prerecorded sound stimuli that match either an AAB pattern (such as ji-ji-li) or an ABB pattern (such as wi-je-je). Once an animal was "habituated" to the pattern, as evidenced by its failure to respond to three consecutive trials, two "test trials" were played and the animal's reaction was noted. According to the description of the experiment, each test trial consisted of three novel syllables (such as ga-ga-ko), with one trial matching and one trial not matching the pattern (AAB or ABB) to which the animal had been habituated.

The *Cognition* paper reported the results of this within-subjects design (at Fig. 2, page B20): subjects did not selectively orient toward the speaker when a test trial matched the pattern of the habituation series (six animals responded and eight did not), whereas animals did selectively orient toward the speaker when a test trial did not match the pattern from the habituation series (twelve animals responded and two did not).

² The videotape is labeled on its face "MARCUS PBACK - GRAMMAR / (AAB vs. ABB) / SECOND ROUND OF TESTING / USING HAB-DISHAB / JAN. 18, 2000 START" and on its spine "MARCUS GRAMMAR PBACK / #2 JAN. 18, 2000." The handwriting appears to be Prof. Hauser's.

B20

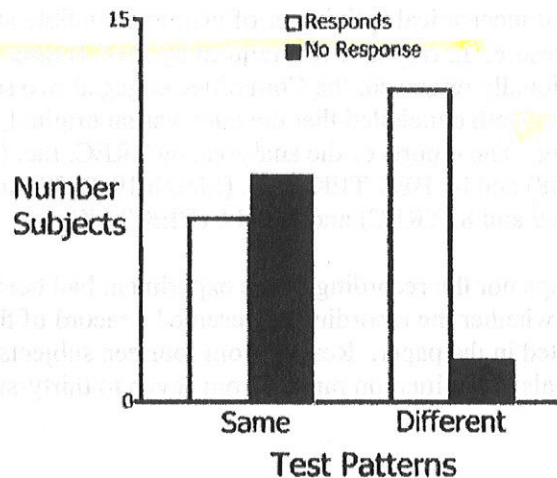
M.D. Hauser et al. / *Cognition* 86 (2002) B15–B22

Fig. 2. Tamarins' responses in the test trials. White bars indicate the number of subjects responding by orienting toward the speaker, while gray bars indicate the number of subjects showing no orienting response.

However, although the VHS recording shows two test trials per animal, neither trial matched the pattern to which an animal was habituated. Only non-matching trials were played. Thus the graph at Fig. 2 does not represent the data on the videotape. The Committee's findings, set forth in detail below, are that Prof. Hauser engaged in research misconduct by knowingly falsifying the research record.

Was the tape altered?

Prof. Hauser proposed two explanations for this discrepancy. First, he suggested that the tape may have been altered, citing differences between the order in which monkeys appeared on the tape and the order noted on the handwritten "log" accompanying the tape (see Appendix Folder II, item 10 for an image of this document).

...it appears to me that the videotape reviewed by the Committee is an edited version, and not the original videotape of the experiments used to generate the *Cognition* 2002 paper. While the overall numbers match up with Figures 1 and 2 in the *Cognition* 2002 paper, the sequence of subjects on the videotape does not match the sequence of subjects in the written transcript (the document found in the sleeve of the videotape box).

(Hauser Response to Committee Allegations 11/27/07, p 3)

I am still not convinced that any mistakes were, in fact made by either myself or (b) (6), (b) (7)(C) I believe there is a real possibility that the videotape may have been doctored. (Hauser letter to (b) (6), (b) (7)(C), 3/6/08, pp. 1-2)

The assertion that the tape had been doctored is discussed at length in "Hauser Response to Committee Allegations 11/27/07," pages 56-64.

Careful examination of the tape by Committee staff revealed no evidence of tampering with the cassette itself, nor of mechanical splicing or of visible or audible artifacts consistent with editing or erasure. In response to a request by Prof. Hauser to have the tape and recordings professionally reviewed, the Committee engaged two separate firms to conduct forensic analyses. Both concluded that the tape was an original, intact, unaltered, unedited recording. The reports of the analyses, by TREC, Inc. (December 8, 2008 and December 19, 2008) and by BEK TEK LLC, (March 10, 2009) are included in Appendix Folder II as items 9 and 8 (TREC) and 7 and 6 (BEK TEK).

Since neither the physical tape nor the recording of the experiment had been tampered with, the next question was whether the recording represented a record of the actual experimental subjects reported in the paper. Results from fourteen subjects were reported. The number of trials to habituation ranged from seven to thirty-six (*Cognition* 2002, page B19, figure 1)

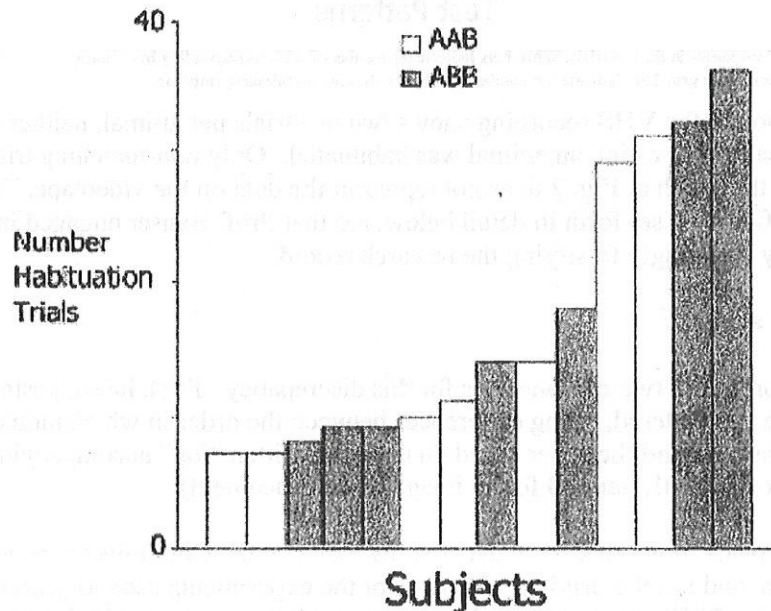


Fig. 1. Number of trials to habituation for subjects tested with either AAB or ABB.

The number of trials and type of trial (AAB or ABB) match the number and type of trial recorded on the videotape, with two exceptions. Only one subject on the tape had seven trials to habituation (subject Irven DeVore, initials ID). Figure 1 reports that two subjects had seven trials to habituation. A spreadsheet found in the Data/Marcus/Raw Data.Results folder on Prof. Hauser's own laptop, and also on Kerouac (a portable hard drive used by Prof. Hauser) shows numbers of trials to habituation that correspond exactly to Figure 1, above; these files also indicate the identities of the monkeys reported. Monkey Irven DeVore appears twice on the spreadsheets, once as ID and once as id. In

both instances the spreadsheets indicate that subject DeVore was habituated to AAB after seven trials and then responded “no” to a matching test stimulus and “yes” to a non-matching test. In addition to the duplicate entries for DeVore, there is one other discrepancy: on the spreadsheets JG is reported with nine trials to habituation, whereas on the tape she had only six. (However, the tape includes a session with RW, not reported in the paper, with nine trials to habituation, and JG was very “jumpy” in her session, which the paper describes as a criterion for rerunning a subject, so the RW data may have substituted for the JG data, or the 6 may have been mis-transcribed as a 9.) These are minor points that the Committee finds do not impugn the authenticity of the tape, and Prof. Hauser concurs that the tape is genuine. (“...the Committee concludes, this tape corresponds well to some of the significant aspects of the experiment reported in the paper. I agree,” and “I agree that the tape provided is most likely the original.” Hauser response, 5/12/09. p 15.)

Were the stimuli loaded incorrectly?

Prof. Hauser’s second, alternate explanation for the discrepancy between the tape and the results reported at Figure 2 is that some test trials could have been incorrectly loaded into the program (a “HyperCard stack”) that generated the stimuli played during the experiment, through inadvertence, and at no time during the setup, running, or analysis of the experiment did anyone discover this error. He provides details about how this could have happened in his 2/8/08 response, starting at page 54; in his 12/10/08 response at page 10 of tab 9; in his 5/12/09 response to the Investigating Committee, pp. 16-20; and in his 5/27/09 letter to the Investigating Committee.

The Committee has carefully considered these explanations, but finds that they are not supported by the written and recorded evidence, and are in fact logically impossible.

In order to understand Prof. Hauser’s assertions, and the Committee’s determinations, it is necessary to understand how the HyperCard stack operates.

The HyperCard stack was written by (b) (6), (b) (7)(C)

The stack was a tool designed to be used in many types of playback experiments, including hab-dishab experiments such as the one reported in the *Cognition* article. A copy of the instructions for use of the stack, written by Fitch, is included in Appendix Folder II, as item 2. When the stack is opened it presents a window that allows the user to name the experiment (this name is used as the title of the output log file generated by the stack), to specify a folder containing the sound files that will be used to habituate the animals (the list), and to load “test” files to be played after habituation. The HyperCard stack displays the name of each sound file loaded into the stack. When it is time to run the study the experimenter presses the “Start” button and then cycles through the list of habituation stimuli by pressing the “Next” button at appropriate intervals. If additional habituation stimuli are needed, the list can auto-randomize and resume. Once the animal is habituated (in this experiment, defined as

failing to respond to three stimuli in a row), the "test" stimuli are played; depending on the protocol for the particular subject, the experimenter presses either (T1) or (T2) for the first test trial, and then the other button for the other test trial. The program creates a record of all stimuli played (both habituation and test trials) in the form of a log file. The log file can include the "online" coding of the experimenter. The title of the log file comes from the field in the upper left corner of the HyperCard window. A stack is prepared once at the beginning of an experiment and then used throughout the experiment. In a study such as the Marcus *Cognition* study, two stacks would have been required, one with AAB habituation materials and one with ABB habituation materials.

Every audio file of Marcus test trial stimuli that we have found on computer disks is correctly labeled. On Prof. Hauser's laptop, for instance, in the Data/Marcus folder, is a folder labeled "test trials" and within that folder are folders labeled "AAB test" (containing two audio files, both of the AAB pattern and correctly labeled, one named "ga ga ko" and one named "ho ho ba"), and "ABB test" (containing two audio files, both of the ABB pattern and correctly labeled, one named "ba ho ho" and one named "ko ga ga"). The creation dates for those files are all January 22, 1999, well prior to the running of the subjects reported in the paper. If test files were incorrectly loaded, the person setting up the experiment would have to have selected both test files from the same folder, rather than selecting one file from the AAB folder and one file from the ABB folder. Further, since the test file filenames correctly indicate their contents, if they were directly loaded into the HyperCard program this error would be apparent at the time the files were selected, when the files were listed in the HyperCard window, and in the text of the log file created during playback.

The (b) (6), (b) (7)(C) instructions indicate that both the habituation ("list") files and the test trials should be auditioned prior to beginning the experiment.

6. Test all the files first. To test sounds in the "List" just double-click on the file name (make sure List Lock is on, or this won't work). Test the "test" stimuli by clicking on the large buttons to the right of the field.

Thus, regardless of how the test files were named, if they were tested per the (b) (6), (b) (7)(C) instructions, it would be apparent whether they were the correct files to be used in the experiment.

The Committee finds that it is highly unlikely that the stimuli could have been loaded incorrectly without the error being discovered by the experimenter(s).

Were the test files renamed?

Prof. Hauser states that

At some point in the history of our running these experiments, we decided that it would be wise, in some cases, for the test trials to be renamed from explicit file names that revealed the precise material to numbers. Although an experimenter

would, of course, know whether the file represented a match or mismatch, its contents would not be revealed until after the experiments. (Hauser statement 5/09/2009, page 17)

To investigate the assertion that files were renamed, Committee staff reviewed every log file from hab/dishab studies that could be found on the Hauser lab disks. Creation dates of such files cover a period of twenty months, from April of 2000 (just after the last subjects in the *Cognition* study were run) to November of 2001. There are fifty-six such files; in no case were the test trials renamed. That is, all files retain the original filenames that indicate the syllables played in the test trials (see Appendix Folder II, item 1).

The instructions written by (b)(6)(b)(7)(C) do not indicate that files are to be renamed. Prof. Hauser states that it was decided that "in some cases" test trials would be renamed; he does not explain why the *Cognition* study would have been singled out for this treatment, and the Committee has not seen anything to distinguish it from every other hab-dishab study where test files were not renamed. If the test files used in the *Cognition* study were renamed, the decision to rename would have to have been made prior to the *Cognition* study, and then immediately reversed for the next fifty-six trials. The Committee finds that all available evidence indicates that the test trials were not renamed.

Could the reported data have resulted from mislabeling of test trial files?

Having acknowledged that the tape is genuine, and despite the analyses above, Prof. Hauser maintains that the experiment must have been run with some test files mislabeled: specifically, two of the four "different" (i.e., non-matching) test files must have been incorrectly labeled as "same" (i.e., matching) files. The design of the experiment, and the comments made by (b)(6)(b)(7)(C) during the running of his subjects, contradict this assertion.

The experimental materials were loaded into the HyperCard program once and then used for the entire experiment.

Once these stacks were in place and set up, they would be used to run all of the monkeys in the colony for this particular experiment. That is, it was only necessary to load up the relevant files once at the beginning of the experiment, and then use the different HyperCard stacks to run the subjects. (Hauser statement 5/12/09, page 16)

Thus, regardless of whether the test files were correctly labeled or mislabeled, their position and imputed structure (AAB or ABB) would have been consistent throughout the experiment. In addition, there are the following facts:

- The pattern each monkey heard during habituation (from the spreadsheets on Prof. Hauser's computers and from the videotape of the experiment).
- How each monkey reacted to a purportedly "same" test stimulus, that is, one that matched the habituation pattern, and to a "different" test stimulus (from the

- spreadsheets).
- The actual patterns of syllables in the test trials played to each monkey (from the videotape of the experiment).
 - The actual responses of the monkeys to the test stimuli (from the videotape, and from present day coding by Prof. Hauser and (b) (6), (b) (7)(C)).

With this information it is possible to determine whether the data represented in the spreadsheets and in the graph at Figure 2 in the *Cognition* article could have arisen from any loading into HyperCard of correctly-labeled and mislabeled stimuli.

Subject SP (Pinker). Per the spreadsheets and per the videotape he was habituated to AAB after 14 trials. He then was played two test trials. To the first test, ba-ho-ho, his response was "yes" (per Hauser and (b) (6), (b) (7)(C) coding) and to the second test, ko-ga-ga, his response was "no."

According to the spreadsheet, the response for SP was "yes" to the "different" trial and "no" to the "same" trial. Since SP was habituated to AAB, this means that the pattern for the "different" trial would have been ABB. Thus, since his response to ba-ho-ho was "yes," ba-ho-ho (pattern ABB) would have been correctly labeled as a "different" trial. If Prof. Hauser's assertion is correct, that the experiment was designed with both matching and non-matching stimuli and that some stimuli were inadvertently mislabeled, it must therefore be the case that the test trial ko-ga-ga was mislabeled as a "same" (or matching) pattern, i.e., AAB.

Subject ID (DeVore). Per the spreadsheet and per the videotape he was habituated to AAB after 7 trials. He then was played two test trials. To the first test, ko-ga-ga, his response was "yes" (per Hauser and (b) (6), (b) (7)(C) coding) and to the second test, ba-ho-ho, his response was "no."

According to the spreadsheet, the response for ID was "yes" to the "different" trial and "no" to the "same" trial. Since ID was habituated to AAB, this means that the pattern for a "different" trial would have been ABB. Thus since his response to ko-ga-ga was "yes," ko-ga-ga (pattern ABB) would have been correctly labeled as a "different" pattern and ba-ho-ho, to which he responded "no," would have been mislabeled as a "same" pattern, i.e., AAB. But this contradicts the labeling in the SP experiment, above.

<i>Subject</i>	SP	ID
<i>Habituation pattern</i>	AAB	AAB
<i>1st test trial played</i>	ba-ho-ho	ko-ga-ga
<i>2nd test trial played</i>	ko-ga-ga	ba-ho-ho
<i>Reaction to 1st test played</i>	Y	Y
<i>Reaction to 2nd test played</i>	N	N
<i>Code for "same"</i>	N	N
<i>Code for "diff"</i>	Y	Y
<i>Inferred pattern of 1st test</i>	Different—ABB	Different—ABB
<i>Inferred pattern of 2nd test</i>	Same—AAB	Same—AAB
<i>Labeling of ba-ho-ho</i>	ABB (correct)	AAB (misabeled)
<i>Labeling of ko-ga-ga</i>	AAB (misabeled)	ABB (correct)

In an attempt to discredit the above analysis, Prof. Hauser, in his 5/27/09 response, describes a scenario in which *four* HyperCard stacks could have been used to run a Marcus-type experiment, instead of two, with two stacks for each condition (habituation with the first trial as a match vs. habituation with the first trial as a mismatch). Prof. Hauser suggests that the individual who set up the experiment created two stacks for each condition, and by mistake loaded test files only from a single folder into both "match first" and "mismatch first" stacks. Thus one stack to be used for animals habituated to AAB had an allegedly-matching test trial in the T1 position, and one had a mismatching test trial in the T1 position, as follows:

Condition	Hab to AAB	Hab to AAB
1 st trial	Match first	Mismatch first
Actual trials loaded into HyperCard	T1: AAB: ba-ho-ho (misabeled as match) T2: ABB: ko-ga-ga (correct label, mismatch)	T1: ABB: ko-ga-ga (correct label, mismatch) T2: AAB: ba-ho-ho (misabeled as match)

However, even this unlikely scenario does not withstand scrutiny. The videotape shows that ko-ga-ga was the first test trial *played* to ID and (b)(6)(b)(7)(C) oral comments indicate that ko-ga-ga was loaded into the second *position* in the HyperCard stack (b)(6)(b)(7)(C) says of DeVore "We're doing **AAB with ABB test with the second stimulus first**"). Thus the arrangement of trials in the stack played to ID would have been as shown in the first column above. This is directly contradicted by analysis of the results from subject SP. Recall that the first test played to SP (habituated to AAB) was ba-ho-ho, that SP's reaction to that test was "yes," and that the spreadsheet said that he reacted "yes" to a "different" (i.e., ABB, or "mismatch") stimulus. This would mean that ba-ho-ho was correctly labeled as ABB ("mismatch") for subject SP. Yet the above arrangements (per Prof. Hauser's suggested explanation) show ba-ho-ho to be "misabeled as match." Given the coding of subjects' responses, ba-ho-ho would therefore have to have been incorrectly labeled as "match" (AAB) in a "match first" stack but correctly labeled as "mismatch" (ABB) in a "mismatch first" stack. Likewise, ko-ga-ga would have to have been incorrectly labeled as a "match" (AAB) in a "mismatch first" stack but correctly labeled as a "mismatch" (ABB) in a "match first" stack. The Committee finds that such a compound error is simply not credible.

Nor would it have been likely, or necessary, to create four separate HyperCard stacks to run the experiment described in the *Cognition* paper. Indeed, the "four-stack" hypothesis is flatly contradicted by oral remarks on the videotape from both Prof. Hauser and (b)(6)(b)(7)(C) as they introduce each experimental session. The remarks make it clear that the order of test trials within the stacks themselves are being varied, rather than using multiple stacks to vary the trial order, and the remarks also imply that *only* mismatching test trials are being played after habituation.

Hauser, testing Spelke: audio is incomplete at introduction, but experimenter states "...**BB condition, tested with AAB.**"

(b)(6)(b)(7)(C) testing Dennett: "This is Dennett with **ABB habituation and AAB test** and let's see, 'ga-ga-ko' first, whichever one that was."

(b)(6)(b)(7)(C) testing Bloom (first Bloom test, data not included in paper): This is Bloom with the Marcus speech materials with **AAB habituation set ABB test** with the first one first. I think it...I left it out on my desk...I don't remember, maybe 'ba-ho-ho'" [his memory was correct]

(b)(6)(b)(7)(C) testing DeVore: "This is DeVore with Marcus-Williams. We're doing **AAB with ABB test** with the second stimulus first."

(b)(6)(b)(7)(C) testing Wrangham: "This is Wrangham with, uh, Marcus Williams stimuli, **CBB [sic] habituation, AAB test stimuli** with the second one first.

(b)(6)(b)(7)(C) testing Bloom (second Bloom test, data not included in paper): "This is Bloom with oh, **AAB habituation, ABB test**, uh, oh, I guess I'll go with the first one first."

(b) (6) (b) (7) (C) testing Goodall: "This is Goodall with ABB for the habituation and AAB for the test. First one first."

(b) (6) (b) (7) (C) testing Wynn: "This is... Wynn with, uh, Marcus trial ABB to AAB and we'll play the second one first."

(b) (6) (b) (7) (C) references to the "first one first" and "second one first" indicate that the test stimuli for the AAB habituation condition, as loaded into one HyperCard stack, were "ba-ho-ho" (first) and "ko-ga-ga" (second), and test stimuli for the ABB habituation condition, as loaded into the other HyperCard stack, were "ga-ga-ko" (first) and "ho-ho-ba" (second). His oral descriptions of which trials will be played, and in which order, correspond exactly to the test trials that were actually played during the experiment. All comments of both Hauser and (b) (6) (b) (7) (C) are entirely consistent with only non-matching test trials being played. The Committee finds that there is no evidence to support the contention that the experiment was run with mislabeled files, or that the experiment involved four HyperCard stacks rather than two.

Could the experiment have been run and data have been analyzed without discovering that only non-matching test files were used?

Despite the many flaws inherent in the mislabeled test trials scenarios, the Committee also considered Prof. Hauser's additional assertion that the experiment could have been run and the data coded and analyzed without anybody discovering that only non-matching test trials were played.

As each animal is run, the HyperCard stack controlling the session creates a log file. Log files automatically include date and time of run, filename(s) of habituation trials, filename(s) of test trials, and information about timing (inter-stimulus interval). The log file might also contain information about the subject or type of experiment (from the title window). According to the *Cognition* article, all trials were scored online, so the log file would also include the experimenter's real-time code for the animal's reaction. The log file is a tab-delimited text file.

If the test files were not renamed (see discussion above), the log file would identify the audio contents of each test file, as below (these log files are from a vervet hab/dishab experiment run a few weeks after the last session on the Marcus tape):

Vervet Playbacks 4/10/00 10:36 AM

#	file	totalTime	ISI	trialType	resp...
1	wi wi li	2.1	2.1	H	Y 1.65
2	ji ji je	28.35	26.25	H	Y 1.57
3	ji ji li	54.73	26.37	H	Y 1.7
4	de de di	79.55	24.82	H	Y 1.8
5	de de li	113.83	34.27	H	N 4
6	ji ji we	160.45	46.62	H	? 16.27
7	ji ji di	181.07	20.6	H	N 4

CONFIDENTIALReport of Investigating Committee
8 January 2010

8	le le di	202.98	21.92	H	N	1.62				
9	Pongo:Desktop Folder:MarcusPB:Test files:ABB Test:ba ho ho	248.33	45.33	T1	N	2.98				
10	Pongo:Desktop Folder:MarcusPB:LEALARM.AIF	268.4	20.07	T3	Y	1.27				

Vervet Playbacks 4/12/00 10:09 AM

#	file	totalTime	ISI	trialType	resp...					
1	wi wi je	1.93	1.93	H	Y	0.65				
2	le le li	20.45	18.52	H	Y	1.52				
3	ji ji di	39.68	19.23	H	Y	1.68				
4	ji ji li	58.65	18.95	H	Y	6.65				
5	de de je	79.3	20.65	H	N	3.73				
6	wi wi li	100.93	21.63	H	N	2.88				
7	ji ji je	127.73	26.8	H	N	4.7				
8	Pongo:Desktop Folder:MarcusPB:Test files:ABB Test:ko ga ga	165.35	37.58	T2	N	4.7				
9	Pongo:Desktop Folder:MarcusPB:LEALARM.AIF	184.2	18.85	T3	Y	3.17				

In these examples, each animal is habituated (three consecutive "No" responses) after a series of trials involving an AAB pattern, and then one test trial of an ABB pattern is run. The log files include name of the test trial files (emphasis added). The HyperCard stack displays the entire pathname of the test trial when the test trial is not in the same folder as the habituation files.

The *Cognition* article describes the coding procedure:

All experiments were videotaped. Although we scored the trials on-line, we re-scored the last three habituation trials and the two test trials by digitizing each trial, and scoring the response blind to condition (see Hauser et al., 2001). Furthermore, and following the procedure used in all other playbacks on tamarins, two experimenters independently scored 20 trials and obtained high inter-observer reliabilities ($r = 0.89$). In these experiments, the on-line scoring for all habituation trials precisely matched those scored blind and thus, we did not have to rerun any sessions. Only five test trials were scored differently on-line and off-line, and we used the off-line response in our analyses. [*Cognition*, p B18]

For this coding process not to have revealed that only mismatched test trials were played, all of the following must have occurred.

1. *The test trials were renamed, incorrectly, before being loaded into the HyperCard program.* But there is no evidence of renaming of any test trials in any hab/dishab experiment, over a period of almost two years.
2. *Contrary to the written instructions accompanying the HyperCard stack, the test trials were not "tested" after being loaded into the program.* If they had been tested, the fact that both test trials were of the same pattern would have been obvious.
3. *The log files output from the HyperCard program were never reviewed.* If the files had been reviewed, it would have been apparent that only mismatched trials were played, since the filename of the stimulus is included in the log file. The description of the coding

procedure indicates that the online and offline scoring of the trials were compared, so the log files *must* have been reviewed.

Even if the files had been renamed before being loaded into the HyperCard stack, the code would have to be revealed at some point to assign conditions, as Prof. Hauser indicates in his 5/9/09 statement.

At some point in the history of our running these experiments, we decided that it would be wise, in some cases, for the test trials to be renamed from explicit file names that revealed the precise material to numbers. Although an experimenter would, of course, know whether the file represented a match or mismatch, **its contents would not be revealed until after the experiments.** (page 17, emphasis added)


4. *The video recordings were digitized without listening to the audio track.* This would have been unusually difficult, since the number of trials to habituation varies from seven to thirty-six and only nine of fourteen subjects were played a post-test-trial screech.

The Committee finds that the overwhelming weight of evidence contradicts the assertion that the experiment could have been run without knowledge that the test trials were only mismatched stimuli.

Might the study originally have been designed with only non-matching test trials?

Neither the written sheet that accompanied the videotape nor the videotape itself contradicts this possibility. The handwritten notes accompanying the videotape all indicate one pattern as "hab" and only a non-matching pattern as "test." The oral remarks by the experimenters imply that only non-matching test trials were played. Prof. Hauser has stated repeatedly that running the experiment with only dishabituation test trials would "make no sense," but experimental log files output from the Fitch HyperCard stack (above) show two other sessions where test subjects were played only non-matching test trials. Indeed, Prof. Hauser himself describes a study where subjects did not receive both matching and non-matching trials in the same session:

For example, in the Ramus et al., 2000, Science work, we presented each subject with a single test trial, half the subjects starting with an inconsistent stimulus (relative to the habituation material) and half starting with a novel string. Ever since the completion of this work, we have shifted to a procedure in which subjects are given multiple test trials, half of which are consistent and half inconsistent, and counter-balancing for which class of stimuli we present first. (Hauser response, 2/8/08, p 52)

 also acknowledged the possibility of a design involving only non-matching trials:

I mean I'm trying now to think ... you just got me thinking about whether in any of the infant work that I've done where we played multiple test trials, and ... I

feel like it's not unheard of, is my gut feeling, though I'm sort of trying to dig because I've run habituation/dishabituation with infants too. And we did run multiple test trials with looking time because you actually ... I mean with looking it's a little bit different than ... I mean the one difference between the way we scored the monkeys and the way we score infant looking time studies are infants pretty much respond to everything, and so you just look at how often ... what's the length or magnitude as measured by the length of their response. Whereas the monkeys, they respond differentially, so you look at whether they respond or not. So in the infant work you definitely ... having multiple test trials is not unheard of because you may take the magnitude of response to this thing from another category across a couple of trials, and then compare back with the magnitude of response to the habituating stimuli. Now I guess that's initially why I never really thought twice about why -- there are two test trials here. There are designs out there that use that type of thing. Whether we just imported it wholesale to a monkey experiment, it's possible. Given the fact that they respond differentially, I don't know what the pattern of results would mean if they didn't respond to one and then responded to the next one--it took them a little while to realize they're hearing something different? It's not entirely clear to me. But it's also not clear to me that you could run an experiment where you would have them dishabituate to a novel category and then play them something from the initial category and expect no response because once their response is rebounded, then you're at a fundamentally different mindset so to speak, to use the term very loosely, than you were after three habituating trials. So it's not clear to me what the expectation ... to mix the trials, it's not clear to me what the expectation is there either. Depending on how you mix them. (b) (6) (b) (7)(C) interview, 5/8/09]

There is nothing to contradict the conclusion that the sessions on this videotape were intentionally designed with only non-matching test trials. This could have been done to explore whether there was a "first trial effect" (as discussed by Prof. Hauser in a document presented to the Committee on December 10, 2008, at tab 10). The more likely explanation is that this tape represents one leg of a study constructed as the Ramus et al. study referenced above, and where subjects would be tested in separate sessions using matching trials. The Committee notes that this interpretation is also consistent with the labels on the videotape ("**SECOND ROUND OF TESTING**" and "**MARCUS GRAMMAR PBACK #2**" on the tape) and the heading on the handwritten log sheet ("**MARCUS GRAMMAR EXPT—2ND ROUND/HABITUATE TO MULTIPLE EXEMP OF ABB OR ABB, THEN TEST**"). Yet the matching trial sessions were either not run, or their data were not included in the paper.

In conclusion, the Committee finds that the videotape represents the experiment reported in the *Cognition* 2002 paper, that there is no "mislabeling" explanation that could account for the results reported (whether two or four HyperCard stacks were used to run the experiment), that the overwhelming weight of evidence indicates that the presence of only non-matching stimuli must have been known, and that the results of the experiment were therefore knowingly and falsely reported by Prof. Hauser.

§93.313 (f) (3) This research received PHS support. The New England Primate Research Center (P51RR00168-37) provided support for the tamarin colony. Prof. Hauser received support from an (b) (6), (b) (7)(C) (cited in the published article, but grant number is unknown).

§93.313 (f) (4) The *Cognition* article will be retracted by agreement of Prof. Hauser and his coauthors. Draft language for the retraction has been proposed.

§93.313 (f) (5) The Committee finds that Prof. Hauser is the person responsible for the misconduct.

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

The Committee is unaware of any pending applications by Prof. Hauser for federal support.

(b) (6), (b) (7)(C) Syllable and Segment (Syl & Seg)

The allegations of research misconduct in connection with this project relate to the coding, statistical analysis, and reporting of the results of two parts of an experiment carried out in the Hauser Lab. The Committee's findings, set forth in detail below, are that Prof. Hauser did engage in research misconduct by intentionally falsifying the research record. Prof. Hauser supervised (b) (6), (b) (7)(C)

Although (b) (6), (b) (7)(C)

The project was a collaboration with (b) (6), (b) (7)(C)

to compare the performance of infants in their lab to the performance of tamarin monkeys in Prof. Hauser's lab, in the statistical task of segmenting a speech stream of an artificial language.

Although the work in question was carried out when Prof. Hauser was on (b) (6), (b) (7)(C) he continued to be involved. He communicated with (b) (6), (b) (7)(C) on a regular basis, and he frequently copied (b) (6), (b) (7)(C) on email correspondence.

Three experiments were run, each using two artificial languages, "A" and "B": AND, VERSUS, and NOT. The AND experiment was run first (b) (6), (b) (7)(C) ran the AND Language A trial (AND LangA) in question on December 15, 2004, and the AND Language B (AND LangB) trial on January 15, 2005 (there had been earlier attempts to run these trials, but those were superceded by the runs in question). No questions have been raised about the running and recording of the trials. In order to evaluate the performance of the subjects, video of the experiments is transferred to digitized video clips of each individual trial. These clips are then scored ("coded").

(b) (6), (b) (7)(C) coded the clips and sent their results to Prof. Hauser (b) (6), (b) (7)(C), in the form of Excel spreadsheets. For the sake of clarity the sequence of events with respect to the AND LangA and AND LangB trials will be discussed separately. Details differ, but the same thing happened in each case (b) (6), (b) (7)(C) had a set of results that was not statistically significant, that Prof. Hauser then changed in a manner that was opaque from the perspective of (b) (6), (b) (7)(C), so that the results then passed the threshold of statistical significance. One reason that (b) (6), (b) (7)(C) did not realize what was happening is that Prof. Hauser ran the statistical analysis on these results, and it was not made clear, when he told them he was making changes or correcting errors, whether he was referring to the statistical analysis or the underlying coding. He provided the statistics, and (b) (6), (b) (7)(C), and collaborator (b) (6), (b) (7)(C) relied on them.

The problem was discovered ten months later, when (b) (6), (b) (7)(C) ran statistics on the AND LangB results that she and (b) (6), (b) (7)(C) had sent to Prof. Hauser for review (labeled with their initials, TO RO), not knowing that the result file that Prof. Hauser used to run the statistics that everyone had was substantially different from theirs because of the changes he had made but not explained.

Prof. Hauser's position, set out in his written response and reaffirmed at his interview, is that the "only alterations I would have made to the initial files [provided by (b) (6), (b) (7)(C)] were in terms of reformatting so that I could import an Excel file into the statistical package Statview." Response, p.44. He goes on to say: "Intentionally changing data points would have served no purpose given that the original video tapes and data files were stored in the lab, and thus could be checked at any time." Response, p.44. The answer to this contention is evident from the facts of this case: it was only because (b) (6), (b) (7)(C) decided to run the statistics on the old TO RO file that the changes were discovered. Otherwise, no one would have had reason to second-guess the coding and painstakingly reconstruct it by reviewing video clips and codes of trials to see whether the action of the subjects was accurately recorded in the Excel files. Furthermore, due to the nature of the statistical analysis used in these experiments, reversing a relatively small number of data points can generate substantial changes in results, so it is unlikely that looking at the result files by themselves would lead to detection of changes. As explained below for example, AND LangB's original p value of 0.073 changed to 0.047 when just 8 of 88 trial results were changed in the direction of significance.

Regarding the storage of research records in the lab, there are two additional issues. Prof. Hauser asserts that, with the exception of a folder with three spreadsheets from the period of the recoding in 2006, the Syl Seg data files are nowhere to be found on his or the lab's computers, except under a User file "timo" set up by (b) (6), (b) (7)(C). Response, p. 42, Hauser Interview 5/19/09, p.15-16. The original drives of Prof. Hauser's computers are in the Committee's possession, and he has been provided exact copies of every drive. Both his laptop and his portable hard drive "Kerouac" have a folder labeled "data." Inside those folders, there are subfolders as follows: Tamarins >(b) (6), (b) (7)(C)> Segment_Syllable Experiments. The "Segment_Syllable Experiments" folder contains eight subfolders,

including subfolders for AND, VERSUS and NOT, that contain 34 files. If the "Segment_Syllable Experiments" folder is not on his copies of the laptop or Kerouac drives, then it was removed from them when those drives were in his custody.

During the events in question, in March 2006, a portable hard drive containing all the Syl Seg data did disappear from the lab. Prof. Hauser contends that (b) (6), (b) (7)(C) blamed him for its disappearance, Response p. 43, but (b) (6), (b) (7)(C) makes no such claim. See (b) (6), (b) (7)(C) interview, p. 8-9, where the Committee raised the question of the disappearance of the drive. While the disappearance of a hard drive with research records from the lab is a serious matter, there is insufficient information for the Committee to make any findings as to the manner of the loss.

What follows is a description of how the initial coding of the AND LangA and AND LangB trials was performed by (b) (6), (b) (7)(C) and the data transmitted to Prof. Hauser, who performed the statistical analysis, including determining the p value.

AND Lang A

Shortly after AND LangA was run on December 15, 2004, Prof. Hauser was the first to code it, and he reported a significant result:

hey all,
so i just did the offline coding of AND-Language-A in the booth (i.e., our old test chamber). A few things, some good, some less so. Overall, there was a significantly greater response to part words than to words (Wilcoxon = 2.22, $p < 0.027$). So that is good. Also good is the fact that (b) (6), (b) (7)(C) have greatly improved the video shot so it is far easier to code. Also, the animal's behavior seems generally better. That said, for this session, we got a fairly low overall level of response from subjects. so out of 10 subjects tested, only 7 yielded data. the other three didn't respond to any playbacks. moreover, several of the others responded to only a few trials. so we will clearly need to run this on language B and hopefully get response rates up. i am assuming we will run the other half on NOT-LangB next. Before doing that, we will put every subject in this new test box and see if we can familiarize them a bit more. in any case, perhaps we are back on track.

a good thing (b) (6), (b) (7)(C)
(Hauser to (b) (6), (b) (7)(C) , 12/20/04)

Prof. Hauser (b) (6), (b) (7)(C) . (b) (6), (b) (7)(C) coded AND LangA and sent their results to him (b) (6), (b) (7)(C) and Prof. Hauser conferred, and used an Internet program to go over the video clips of AND LangA together in order to review the coding that (b) (6), (b) (7)(C) had sent. Following this review session, Prof. Hauser sent the revised codes to (b) (6), (b) (7)(C) on February 3, 2005 [TAM Syl AND Seg Lang A 12.15.04] and told (b) (6), (b) (7)(C):

"so with this new analysis we lose the effect. it is $p=.18$. So, if Lang B is also

weak, that sinks it.” (Hauser to (b) (6), (b) (7)(C) 2/3/05)

(b) (6), (b) (7)(C) replied:

“that sucks. So assuming that there is nothing in the B run after we do a proper code do we scrap it and give up on this stuff?” (b) (6), (b) (7)(C) to Hauser, 2/4/05)

Prof. Hauser promptly replied:

(b) (6), (b) (7)(C)

hold the horses. i think i fucked something up on the coding. let me get back to you.

marc

(Hauser to (b) (6), (b) (7)(C) 2/4/05)

Prof. Hauser then got back to (b) (6), (b) (7)(C) later that day:

yes, i fucked up. i was playing some more with the data, looking more carefully at some of the patterns, and in stat view i reversed the pw vs w coding for one animal. so this one they pass at .04. not huge, but they pass. note that two animals didn't respond at all. so, numbers are small here. we will need to do lang b carefully. (Hauser to (b) (6), (b) (7)(C), 2/4/05)

(b) (6), (b) (7)(C) replied by asking to see the codes, so he and (b) (6), (b) (7)(C) could learn from the changes Prof. Hauser made:

Send me the file with the tie breaks we did earlier so I can look at it with (b) (6), (b) (7)(C) when you get a chance. (b) (6), (b) (7)(C) to Hauser, 2/4/05)

Instead of sending the file, Prof. Hauser told (b) (6), (b) (7)(C)

the reason for the recode was that i accidentally trashed the file that had the reworking of our efforts. all i have is the attached, which is the final coding of lang a based on our discussion this morning. (Hauser to (b) (6), (b) (7)(C), 2/4/05)

What Prof. Hauser sent was a single column of coding, with a different file name than the file he had sent before the session with (b) (6), (b) (7)(C)

On May 23, Prof. Hauser sent (b) (6), (b) (7)(C) the final codes [AND_LangA_Final.xls] and the accompanying statistics summary of the same date, stating a p value of 0.01 for AND LangA. AND_LangA_Final.xls had five data points changed from the file sent to (b) (6), (b) (7)(C) on February 3. Four of the five changes were in the direction of significance.

AND LangB

The AND LangB run at issue took place after Prof. Hauser was in (b)(6) on sabbatical (b)(6) (b)(7)(C) coded the clips from the January 18, 2005 run and sent DVDs of the clips and a spreadsheet of their codes. (7)(c)

In the March 17 email accompanying the code spreadsheet [ANDLangB011805_toro.xls], (b)(6) (b)(7)(C) pointed out the column of results where he and (b)(6) (b)(7)(C) had differed and arrived at a negotiated result, so that Prof. Hauser could review those trials:

Note that I am pretty sure that despite the labeling on the column that the tie breaks for this run are not yours but were negotiated between me and (b)(6) (b)(7)(C) unless you remember differently. So if you think it might make a difference you may want to tie break those. (b)(6) (b)(7)(C) to Hauser, 3/17/05)

On March 18 (b)(6) (b)(7)(C) re-sent the spread sheet, and reminded him:

Right, so here is the run from the 18th again [01 18 05]. This is the one you want. The column labeled "tie" has the ones where (b)(6) (b)(7)(C) and I disagreed and negotiated. (b)(6) (b)(7)(C) to Hauser, 3/18/05)

A file called ANDLangB_HalfCol2.Final.xls, is on Prof.'s laptop computer and portable hard drive "Kerouac," in the folder that Prof. Hauser denies having on his computer, see above. Comparing the codes in the column "FINAL Response" on that spreadsheet with the "tie" column on the spreadsheet sent by (b)(6) (b)(7)(C) on March 17, one sees that the codes for six trials are changed, all in the direction of significance. However, statistical significance is not achieved when calculated from these results. The ANDLangB_HalfCol2.Final.xls file on Prof. Hauser's computers was last modified May 18, 2005. Prof. Hauser did not send this spreadsheet to (b)(6) (b)(7)(C)

When Prof. Hauser emailed (b)(6) (b)(7)(C) to confirm the p values he had for AND LangA and AND LangB, (b)(6) (b)(7)(C) responded by pointing out that he had not done any statistical analysis for either of those. (b)(6) (b)(7)(C) to Hauser, 5/17/05) The emails and spreadsheets set forth above confirm that Prof. Hauser, not (b)(6) (b)(7)(C), carried out the statistical analyses.

On May 22, 2005, Prof. Hauser sent the ANDLangB_Final code sheet to (b)(6) (b)(7)(C) without sending any spreadsheet showing the (b)(6) (b)(7)(C), or "negotiated" codes along with his changes (he had changed six of the "negotiated" codes, all in the direction of significance), and without informing (b)(6) (b)(7)(C) that he had changed two codes that (b)(6) (b)(7)(C) had agreed upon:

By the way, here is what I will send to (b)(6) (b)(7)(C) in terms of raw data. This is from lang_B AND. Why don't you send me similar for NOT and VERSUS and I can send on. (Hauser to (b)(6) (b)(7)(C), 5/22/05, with AND_LangB_Final attached)

The ANDLangB_Final codes differ slightly from the codes in the column "FINAL Response" on ANDLangB_HalfCol2.Final.xls. Two additional codes, that (b) (6), (b) (7)(C) had agreed upon, are changed in the direction of significance. With the change of those two codes, the p value goes from $p=0.084$ (column "FINAL Response" on ANDLangB_HalfCol2.Final.xls) to $p=0.047$ (ANDLangB_Final).

As requested by Prof. Hauser, (b) (6), (b) (7)(C) sent him spreadsheets in the same format as ANDLangB_Final for the other experiments. Prof. Hauser corrected and returned the spreadsheets on May 23, 2005, along with a summary of statistics showing p values of 0.01 and 0.047 for AND LangA and AND LangB respectively.

Prof. Hauser reported all the results to (b) (6), (b) (7)(C) and suggested that they begin drafting their paper:

ok, so (b) (6), (b) (7)(C) (b) (6) versus lang b and it is again a failure. the wilcoxon is .32 with $p=.75$. no hint of a success. so that gives us big successes on AND and none on NOT and VERSUS (b) (6), (b) (7)(C) can we put these together with baby data for one paper? ready to roll? (Hauser to (b) (6), (b) (7)(C) , 6/19/05)

Findings of fact regarding the coding and analysis of AND LangA and AND LangB

Based on its review of the email between Prof. Hauser and (b) (6), (b) (7)(C) during the period January, 2005 to May, 2005, and the spreadsheets and data files related to AND LangA and AND LangB as summarized above, the Committee finds that Prof. Hauser did change data points in a manner that altered the outcomes of the statistical analyses of the studies, without explaining to (b) (6), (b) (7)(C) what he had done or why he did it.

With respect to AND LangA, Prof. Hauser provided (b) (6), (b) (7)(C) with a data set and reported that his statistical analysis yielded a p value of 0.18. The date was February 3, 2005. At no point after that did Prof. Hauser tell (b) (6), (b) (7)(C) that he had made further changes to the data set. While the record does not enable the Committee to reconstruct every step of the sequence leading from the announced p value of 0.18 to a claimed p value of 0.01, the Committee finds that Prof. Hauser changed five data points in the data set that was labeled "Final" and sent by him to (b) (6), (b) (7)(C) on May 23, 2005, for forwarding to (b) (6), (b) (7)(C). Four of the five changes were in the direction of significance, and included egregious violations of coding protocol, reversing earlier coding decisions made by (b) (6), (b) (7)(C), and Hauser himself. For example, subject AG, trial 4 (clip 53) and subject JG, trial 7 (clip 46) had both been coded "no" but Prof. Hauser changed them to "yes," a change in the direction of significance. The Committee finds no basis for calling these trials "yes" responses. These changes constitute falsification.

Prof. Hauser's changes to the AND Lang B codes violated coding protocol as well. He reviewed a set of codes submitted by (b) (6), (b) (7)(C) and arrived at a consensus regarding trials where they initially disagreed. Prof. Hauser

changed six of the codes where they had negotiated. In each case Prof. Hauser's change was in the direction of significance. He changed two others where (b) (6), (b) (7)(C) had agreed, also in the direction of significance. These latter two, subject PJ trial 2 (clip 18) and subject PB trial 7 (clip 46) had been coded "yes" by his students, but Prof. Hauser changed them to "no." Taken together, these four changes from LangA and LangB demonstrate that Prof. Hauser was manipulating the trial codes in order to have data sets that would yield statistically significant results. He relabeled the two "yes" trials "no" and the two "no" trials "yes." Looking at the four trials one after the other, two are readily identifiable to the Committee as "yes" trials and two are "no," but Prof. Hauser's recoding has them backwards.

The Committee finds that Prof. Hauser misled (b) (6), (b) (7)(C) about the changes he had made to AND LangB data as well. On January 13, 2006, after (b) (6), (b) (7)(C) had told (b) (6), (b) (7)(C) had run statistics for AND (b) (6), (b) (7)(C) asked Prof. Hauser for assistance in understanding the source of the inconsistencies:

(b) (6), (b) (7)(C) and I did not run any of the final analyses for any of that stuff, we simply sent you and (b) (6), (b) (7)(C) the data files. (remember I tried to run my first analysis ever in late spring and munged it somehow, so you redid everything) Since, we didn't have the original analyses and (b) (6), (b) (7)(C) needed them for her write up, we thought we would just regenerate them from the original data files. When we did this we got the $p = .73$ result on the original AND lang B run. That is from the original file which we have from when we first coded the stuff.
So if you could walk me through exactly how you have been doing these analyses,
and I will walk you through what I am doing and we can see if there are discrepancies, also we can compare the actual data files we are looking at.
(b) (6), (b) (7)(C) to Hauser, 1/13/06)

Prof. Hauser answered by sending the AND LangB Final code file that contained the six changes of (b) (6), (b) (7)(C) "tie breaks" in favor of significance, plus the two changes of codes where (b) (6), (b) (7)(C) agreed, which he had originally sent to (b) (6), (b) (7)(C) on May 22, 2005, see above:

ok, so i just reran the stats on and_B and got $p = 0.047$ so significant. am attaching
the excel which i labeled as final after coding and tie breaking.
(Hauser to (b) (6), (b) (7)(C) 1/13/06)

(b) (6), (b) (7)(C) I promptly replied, asking for the code file containing his and (b) (6), (b) (7)(C) codes as well as the tie break record ("the usual code file format"):

thanks, I will look at this. do you have the original file you got this from, I see that
this is in the format you put it in before you do your stats, but I sent you the usual

code file format, do you have that one still? (b) (6), (b) (7)(C) to Hauser, 1/13/06)

Prof. Hauser answered:

No, this is just your code but reformatted. I threw out the original, but this is it, except without some of the video files [the columns of run times and elapsed times]. (Hauser to (b) (6), (b) (7)(C), emphasis added, 1/13/06)

The Committee finds Prof. Hauser's assertion that AND LangB Final was "just your code but reformatted" to be untrue. It was a response that would tend to discourage further investigation and maintain the status quo: if it had been accepted at face value (b) (6), (b) (7)(C) would have used the data set and the statistics from it to (b) (6), (b) (7)(C), reporting a significant effect. Instead, (b) (6), (b) (7)(C) prepared a spreadsheet that highlighted the eight codes he changed that moved the p value from 0.73 to 0.047, and that contained a separate sheet with their original coding with the negotiated "ties," TAM_AND_LangB_01.18.05_all.xls. (b) (6), (b) (7)(C) I sent it to Prof. Hauser on January 25. Prof. Hauser interprets his reply:

this email, sent on 1/25/06 to (b) (6), (b) (7)(C) shows that I was keen on recoding the work, as opposed to ignoring the issues: "well, at this point i give up. there have been so many errors, i don't know what to say except that you and (b) (6), (b) (7)(C) should recode every single trial from every single run and get a final word on the experiment. I have never seen so many errors and this is really disappointing. see you in the morning." These are clearly not the words of a researcher trying to surreptitiously alter data.
Response, p. 44

These may not be the words of someone trying to alter data, but they could certainly be the words of someone who had previously altered data: having been confronted with a red highlighted spreadsheet showing previous alterations, it made more sense to proclaim disappointment about "errors" and suggest recoding everything than, for example, sitting down to compare data sets to see how the "errors" occurred.

Early the next morning, Prof. Hauser sought to regain control of the AND coding. At his interview Prof. Hauser denied that the AND trial was pivotal (Hauser Interview 5/12/19, p.10), but it was, as shown by his exchange with (b) (6), (b) (7)(C) when he reported the bad result for AND A ("So, if Lang B is also weak, that sinks it." (Hauser to (b) (6), (b) (7)(C) 2/3/05)), and subsequent email to (b) (6), (b) (7)(C), set forth below.

actually, what i would like to do is completely recode AND myself. why don't you burn the CD with all the trials for AND_A and AND_B and i will recode when i get back. why don't you and (b) (6), (b) (7)(C) completely recode NOT and VERSUS. (Hauser to (b) (6) and (b) (7)(C) 1/26/06).

Since (b) (6), (b) (7)(C) was based on Syl Seg. (b) (6), (b) (7)(C) decided to recode AND herself, with (b) (6), (b) (7)(C). (b) (6), (b) (7)(C) also told Prof. Hauser (b) (6), (b) (7)(C) would like to receive his recoding of

AND. When Prof. Hauser heard that (b) (6), (b) (7)(C) had recoded AND and found that the results were not significant, he made a "kind of radical" suggestion, that (b) (6), (b) (7)(C)

a few things. first, it wasn't clear to me what we need to do on AND. our goal was to recode all of the data due to confusions, have two coders and one tie breaker. has this be done for all of the runs? second, if i understand your analyses on AND, and if this is complete, none of our studies showed a significant effect. if that is the case, then, i think you may want to rethink (b) (6), (b) (7)(C). i realize this is kind of radical, but since we have thought that we had this result, and i really don't understand what happened on coding, you basically have no results. had we caught this error earlier, assuming that we have a null result, we could have run other experiments for (b) (6), (b) (7)(C). but as it stands, if these analyses are correct, we don't have any effects, and thus, no results. with a failure on AND, running NOT and VERSUS didn't make any sense. ...so, the first question is whether we have truly done a complete and clean recoding of all this. if not, we should. if so, and these analyses are correct, i would recommend (b) (6), (b) (7)(C). i don't think it will do you any harm at all (b) (6), (b) (7)(C) (Hauser to (b) (6), (b) (7)(C) 2/22/06 emphasis added)

In effect, Prof. Hauser was seeking to prevent publication of a paper reporting the failure of AND. (b) (6), (b) (7)(C) replied that she wanted to discuss how she could (b) (6), (b) (7)(C) despite the failure on AND, and Prof. Hauser repeated and expanded his argument in opposition:

... (b) (6), (b) (7)(C) if there are really no results, (b) (6), (b) (7)(C) as i said, had we picked up on the failure with AND, we wouldn't have run NOT and VERSUS. The failure on AND could be do [sic] to many reasons, but probably uninteresting as we know that saffran original repeated beautifully, and AND is functionally the same statistically, so failure is either their boredom, cleaner stim which fail to trigger detection, motivation, etc. That just isn't interesting to anyone. let me be clear: i don't think this is your fault. this is our collective fault for not having detected errors... do call at home as i am more than happy to talk. i can imagine that this is upsetting because i know how hard you have worked. but let's talk marc (Hauser to (b) (6), (b) (7)(C) 2/22/06)

Faced with Prof. Hauser's resistance (b) (6), (b) (7)(C), (b) (6), (b) (7)(C) on the evening of the 22nd, and reported (b) (6), (b) (7)(C) meeting the next morning:

i recoded all of the runs. we then used last year's final code as the 'second coder' and (b) (6), (b) (7)(C). my correlation with last year's code was very high. i can get the exact numbers for you on saturday when i get

back from (b)(6). it would be great for you to code the two AND runs though, if you would like, since those are the two questionable runs.
(7)(c)

i met with the head of my department last night to ask him about department policy and honors and what his advice was. he said that even given a negative result, under these circumstances and given how much work i have done, the department is fine with me writing this (b)(6), (b)(7)(C) and he said that the lack of result would not hurt my chances of getting (b)(6), (b)(7)(C), but it is at least a consolation that this is possible. i am going to go ahead and do this. my discussion chapter will focus on why i believe the tamarins failed. which means there is one more thing to do--which is very important for all of us at this juncture i believe. i'd like to rerun saffran. we've twice gotten very significant results on this, and if we get another, it shows that for syl seg it was a problem with the stimuli. if the monkeys fail, then we know they are burnt out on the method....

i would still like to meet monday to see if we can come up with any ideas for ways to make this a little better. (b)(6), (b)(7)(C) to Hauser, 02/23/06, emphasis added)

The highlighted passage of (b)(6), (b)(7)(C) email shows that she had already developed a plan to move forward and see what could be learned from the newly discovered null effect. Prof. Hauser includes his second February 22 email to (b)(6), (b)(7)(C) ("(b)(6), (b)(7)(C)..." above) in his Response, p., 50. His discussion of the letter bears close consideration. He claims that the letter "highlights three important points," the third of which is: "I do not say that she should not or must not (b)(6), (b)(7)(C). While it is true that saying "i am afraid there really isn't a coherent (b)(6), (b)(7)(C). ... That just isn't interesting to anyone." is not an order to (b)(6), (b)(7)(C), those words communicate "should not." Prof. Hauser's parsing of the record is misleading in any event, since this letter must be read in the context of his earlier email the same day, where he *did* say: "...i would recommend (b)(6), (b)(7)(C) i don't think it will do you any harm at all (b)(6), (b)(7)(C) (Hauser to (b)(6), (b)(7)(C) 2/22/06, above).

Based on all the evidence in the record, the Committee finds that Prof. Hauser falsified data in both the AND LangA and AND LangB files so that significant results could be reported and used as the comparison for the NOT and VERSUS experiments.

Ultimately, (b)(6), (b)(7)(C) did carry out the recoding for (b)(6), (b)(7)(C), where the AND LangA and AND LangB results were reported as follows:

In each language and when both are combined, tamarins showed no significant difference in the amount of orientation to partwords and words (Wilcoxon signed rank test; $z(8) = -1.19$, $p = 0.23$ for Language A; $z(9) = -1.37$, $p = 0.17$ for Language B; Combined $z(17) = -1.583$, $p = 0.11$). The lack of a significant difference in responses to partwords and words indicates that the tamarins were not distinguishing between the two, and thus were not successfully segmenting

the speech stream. (b) (6), (b) (7)(C)

Prof. Hauser acknowledges the validity of these results: "yes, i think the current data are OK. i agree. (Hauser to (b) (6), (b) (7)(C) 3/13/06); "when the analyses were rerun...there were no significant effects" (Hauser submission to Committee, 12/10/08 at tab 15).

§93.313 (f) (3) This research received PHS support. NIH/NIDCD sponsored award number 5 R01 DC005863-03. Funds from this award provided support for the tamarin colony and for the work of (b) (6), (b) (7)(C).

§93.313 (f) (4) The Committee is not aware of any publications that need correction or retraction as a consequence of this misconduct.

§93.313 (f) (5) The Committee finds that Prof. Hauser is the person responsible for the misconduct.

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

The Committee is unaware of any pending applications by Prof. Hauser for federal support.

**P vs. N, "Grammatical pattern learning by human infants and monkeys"
(manuscript submitted to various journals)**

Introduction

Prof. Hauser carried out the P vs. N project in collaboration with (b) (6), (b) (7)(C) the University of Wisconsin, Madison. (b) (6), (b) (7)(C) ran experiments using infants as subjects in parallel with Prof. Hauser's experiments with the tamarins in his laboratory. The experiments follow the "grammar expectancy violation" model: subjects are habituated to human voice recordings of artificial "words" that follow a grammatical structure, and then they are exposed to stimuli that either conform to, or violate, the grammatical structure and it is judged whether the subjects respond. (b) (6), (b) (7)(C) was involved in the early stages of the project, and was a listed coauthor on the manuscripts. He had not been involved with the project for over a year when he renewed his attention in early 2006, when concerns about coding in the lab surfaced. (b) (6), (b) (7)(C) ran trials for some of the experiments in the project. (b) (6), (b) (7)(C) ran trials for another experiment. The manner in which the trials were carried out is not at issue.

Manuscripts based on P vs. N have been submitted to journals numerous times. Early versions were based on Experiment 1 and Experiment 2 of P vs. N; Experiment 3 was

carried out later in response to a reviewer's comments and suggestions. No claims of misconduct were made about Experiment 2.

When the protocol for the project was developed, the coding plan was to have two observers code blind to condition, and to have a third observer act as a "tie breaker" where the other two disagreed about how a trial should be coded. The protocol was not followed, in the first instance, for either Experiment 1 or Experiment 3. When irregularities were discovered by (b)(6) (b)(7)(C) after the manuscript had been submitted several times, Experiments 1 and 3 were recoded and the manuscript was revised to reflect the results of the recoding. The Committee's findings, set forth in detail below, are that Prof. Hauser did engage in research misconduct by intentionally falsifying the research record. The claims and findings about Experiments 1 and 3 will be discussed separately.

Experiment 1

It is undisputed that the report of the Experiment 1 trials in manuscripts submitted to *Cognition*, *Science*, and *Nature* was based on statistical analysis of data in spreadsheets that was transferred from handwritten coding sheets found in the Hauser Lab. App. IV, P vs N, item 8, [*P vs N Expt 1 coding sheets.pdf*]. At issue is who carried out the coding, and whether it was carried out blind. It is undisputed that if an observer coded by writing results directly on the coding sheets while viewing video clips, coding would not be blind because the column where codes were entered is adjacent to a column indicating whether the stimulus was grammatical or ungrammatical.

Prof. Hauser admits that a substantial number of the trial codes are in his handwriting. He questioned whether the "no" results indicated by a capital "N" bearing exaggerated serifs were his. See Hauser Response 5/12/09, p. 54, Hauser Interview 5/19/09, p. 20 - 21. Prof. Hauser contends that, whether or not all the handwritten code entries were made by him, he coded the Experiment 1 video clips blind to condition. He claims that he would have created a file that only had a column of randomized numbers representing the subject and trial, not any columns that would have disclosed the condition. He stated that after entering his codes in that file, he would have made the written entries on the coding sheets in question, and then he would have made an Excel file to load the codes in StatView, the statistical analysis software program he used. Hauser Interview 5/19/09, p. 21-22. No file containing the codes and randomized numbers, or subject and trial number without condition information, has been found. Prof. Hauser stated that he would have discarded that, Hauser Interview 5/19/09, p. 21-22.

The Committee finds that Prof. Hauser entered the "no" responses referenced by the "N's" discussed above. The handwriting on the coding sheets is obviously his. It is highly improbable that he would have transcribed some, but not all, of his codes onto these sheets, and then that some other unknown person would have entered additional "no" responses to complete the lists; no such process has been described as lab practice and procedure, and there is no sensible explanation for why that would have happened. Regarding the question of whether Prof. Hauser entered the results on the sheets while

viewing the video clips – i.e., while coding – the Committee finds that he did, so he was not coding blind to condition. The notations on the sheets are extensive, and include incidental observations of occurrences seen in the video clips. Transcribing these from a separate, blinded form onto this data sheet after the fact would have been time consuming and unnecessary. These coding sheets are clearly the work of a person viewing the videos and making the entries.

The Experiment 1 results that Prof. Hauser coded in violation of protocol were reported in manuscripts submitted to *Cognition*, *Science*, and *Nature*. The manuscripts stated that “two blind observers” coded trials and a third coded trials to resolve differences [as a tie breaker]. *Nature* rejected it as written, but suggested that an additional experiment be run (referred to below as “Experiment 3”). [REDACTED] and Prof. Hauser devised and ran Experiment 3, revised the manuscript to incorporate findings about it, and resubmitted the manuscript.

The revised manuscript no longer mentioned “two blind observers” and a third to break ties, but it still stated, as all previous versions had, that “Inter-observer reliabilities ranged from 0.85 to 0.90.” To report values for inter-observer reliability when there was a single coder is to misrepresent that more than one observer coded the trials. When *Nature* rejected the revised version, it was submitted to *PNAS* with the same misrepresentation about inter-observer reliabilities. Accordingly, the Committee finds that Prof. Hauser committed research misconduct by intentionally falsifying the report of the coding procedure he used.

The versions of the manuscript submitted to *Cognition*, *Science*, and *Nature* reported that in the predictive language condition, “16 out of 16 subjects” responded more to the ungrammatical than the grammatical stimuli (that claim was omitted from the *Nature* revision; nothing was said about the proportion of subjects responding more to ungrammatical stimuli).

Prof. Hauser’s handwritten coding sheets App. IV, P vs. N, item 8, [*P vs N Expt 1 coding sheets.pdf*] and the spreadsheets based on them do reflect that there were 16 subjects, but they show that one of the sixteen responded more to grammatical than ungrammatical stimuli, and one responded equally to grammatical and ungrammatical. Hence, reporting “16 out of 16 subjects” was a misrepresentation. When Experiment 1 was finally recoded in accordance with the original protocol (two coders plus tie breaker) the result was that 13 subjects responded more to grammatical and 3 responded equally to grammatical and ungrammatical. App. IV, P vs N, item 10, [REDACTED]_Expt1_Recode_020106copy.xls]

The Committee finds that Prof. Hauser recklessly or intentionally falsified the results of Experiment 1 by reporting in manuscripts submitted to *Cognition*, *Science* and *Nature* that in the predictive language condition, “16 out of 16 subjects” responded more to the ungrammatical than the grammatical stimuli. While Prof. Hauser correctly notes [Hauser 5/12/09 response, page 54] that the result reported was significant regardless of the misrepresentation, that does not absolve him. Misrepresenting the actual experimental data to exaggerate the strength of results is falsification.

Experiment 3

Experiment 3, "Medium Grammar," was designed and carried out as a means of bolstering the manuscript being revised and resubmitted to *Nature*. At issue is whose coding and analysis was reported in the revision that was submitted to *Nature* in September, 2005, and whether that coding was properly conducted.

It was (b) (6), (b) (7)(C) who raised the question of whether the reported Experiment 1 results were based on coding that did not conform with the project's protocol. After that, (b) (6), (b) (7)(C) urged that Experiment 3 be recoded: "...we owe it to ourselves to figure out exactly who coded what, since the same sort of confusion that occurred in Experiment 1 appears to have occurred in Experiment 3." (b) (6), (b) (7)(C) to Hauser, 2/5/06)

In answering that email, Prof. Hauser stated:

< Of course we want the right answers. To think otherwise, (that is, to think that I would want anything else) is pure BS! I resent the implication to be honest. I had thought that (b) (6), (b) (7)(C) had coded expt 3 of saffran. (Hauser to (b) (6), (b) (7)(C), 2/6/06)

Prof. Hauser summarized these events as follows, in his initial response to the Committee on Professional Conduct, before any allegation had been made regarding P vs. N:

After I thought we had completed all of the analyses, and had a nearly completed manuscript, my RA emailed to tell me that he thought that there were possibly some significant errors in coding. In looking at my own email response, it is clear that I was initially defensive and annoyed as I thought we were done and that everything had been checked. I thus challenged the need to recode. (Hauser Response 9/17/07, p.20)

Based on the sequence of events surrounding the coding of Experiment 3 that is set out in the following email and the irregularities that were found in the coding, the Committee finds that Prof. Hauser "challenged the need to recode" in order to avert discovery of the fact that the manuscript submitted to *Nature* misrepresented the coding and results of Experiment 3.

There had been several partial runs of Experiment 3, by (b) (6), (b) (7)(C). These did not yield sufficient results to analyze and report. (b) (6), (b) (7)(C) Prof. Hauser agreed that further trials should be run. Prof. Hauser inquired when he expected that this latest run had taken place:

Hey guys

Any updates on the saffran run. We would like to turn our paper around asap. If there is coding to do, I can do it. Just pass the files on to me.

Thanks (Hauser to (b) (6), (b) (7)(C) 8/22/05)

Prof. Hauser then wrote, subject: saffran rerun:

Since the saffran run is really my work, not yours, if you guys can do the next run (or even just (b) (6), (b) (7)(C) under the RA position role), I will do the coding if you hand off to me. (Hauser to (b) (6), (b) (7)(C), 8/26/05)

He checked in a few days later, subject: run today:

How did it go? Do we have enough now and when might you guys get me the files for coding. I appreciate the help on this! (Hauser to (b) (6), (b) (7)(C), 8/30/05)

They answered that the run went well, and that the "whole colony" was run, and he replied:

Awesome. Can you guys get me clips and excel file soon? I can then code. (Hauser to (b) (6), (b) (7)(C), 8/31/05)

On September 3, (b) (6), (b) (7)(C) told Prof. Hauser that he was going to be out of the lab for the weekend, and had left the DVD of clips being burned in one of the lab computers, where Prof. Hauser could find it:

when I left the other dvd was burning in the g5 machine next to the door of the workshop. I imagine it is still in the drive there. Look at the file names for the particular files as they are prefixed with EUG, EG, HUG, HG, etc. Ask (b) (6), (b) (7)(C) about which monkeys she already ran, although I suggest you try to get a result with just these 20 since she did something funny choosing the monkeys last time. She reran half of A and half of B, under the theory that we could just add that half and half to the existing data -- so it wasn't clear to me which monkeys had been run twice, etc. -- which is why I decided to run the whole colony this time so you would have clean data. Plus this was an Excellent run behavior wise. lots of unpacking to do. hoping to get out of boston tomorrow. (b) (6), (b) (7)(C) to Hauser, 9/3/05)

Prof. Hauser acknowledged that he had found the DVD:

Got it. Yeah, I figured I would look at this run. Perhaps the cleanest I have seen!!! Good luck with leaving....(Hauser to (b) (6), (b) (7)(C), 09/04/05, 5:54 a.m.)

By the end of that very day, Prof. Hauser had coded Experiment 3, analyzed, and reported results to (b) (6), (b) (7)(C) to incorporate in the *Nature* manuscript revision App. IV, P vs N, Item 3, [TAM_LastRun_0830.EDIT.xls]:

ok, so we finished the run. we have an N of 20, and the run was the best i have seen. if you analyze all 8 trials, that is 4 grammatical and 4 ungrammatical, they

fail: Wilcoxon = 1.69, $p = 0.091$.

If you pull out the easy tokens, the ones that violate everything and thus, should be the most novel, you get a success with an $N = 12$ (the N goes down because of some subjects, there were no values as these happen to hit the Bad trials):

Wilcoxon = 2.18, $p = 0.029$. This is good because it shows that even with the simplified rule structure and reduced lexicon (so to speak), they still can't extract the grammar. but because they at least respond more to the easy ungrammatical strings, they are attending and making some kind of discrimination.

Here are the descriptive stats for the entire sample:

Mean/SD/SE response to Grammatical: 26.25/25.47/5.70

Mean/SD/SE response to Ungrammatical: 38.75/29.17/6.52

Here they are for the easy trials:

Mean/SD/SE response to Grammatical: 33.33/40.82/10.54

Mean/SD/SE response to Ungrammatical: 63.33/29.68/7.66

I take it this is all you need to plug and play. perhaps give this a whirl and then send on to me (b) (6) (b) (7) (C) for final comment. (Prof. Hauser to (b) (6) (b) (7) (C) 09/04/05, 10:58 p.m.)

(b) (6) (b) (7) (C) sent the draft with these data included, and Prof. Hauser sent his final comments on September 9, 2005. Then the manuscript was submitted. Since Experiment 3 was only coded by Prof. Hauser, the misrepresentation based on the reference to "inter-observer reliabilities," discussed above, relates to it as well.

The Committee finds that, since Prof. Hauser had told (b) (6) (b) (7) (C) that he would take care of the coding on Experiment 3, and he coded, analyzed, and sent the results to (b) (6) (b) (7) (C) all in one day, his violation of protocol and the misrepresentation of the protocol in the manuscript was intentional. He was in a hurry to resubmit the manuscript with the Experiment 3 results: "...We would like to turn our paper around asap. If there is coding to do, I can do it. Just pass the files on to me." (Hauser to (b) (6) (b) (7) (C), 8/22/05)

The revised manuscript was rejected by *Nature*, and (b) (6) (b) (7) (C) and Prof. Hauser decided to submit it to *PNAS*, after making slight revisions so it conformed to the editorial standards of that journal. The *PNAS* manuscript contained the same misrepresentation regarding "inter-observer reliabilities." *PNAS* reviewers raised questions that (b) (6) (b) (7) (C) Prof. Hauser were in the process of addressing when (b) (6) (b) (7) (C) asked that they hold off on resubmitting:

btw, now that you seem to be digging up errors, were you also suggesting that stats in expt 2 in our paper were not right? if so, could you send me what you have as (b) (6) (b) (7) (C) and i were just about to send off a revision as we think pnas may accept it.

marc

(Hauser to (b)(6)(b)(7)(C), 2/1/06)

I rechecked Experiment 2, phrase structure, and those stats are all correct and everything looks as it should. But I would encourage you to hold off, if only for one day, to give time to recheck experiment 1. I am putting this at the top of my priorities so that we can get it off to PNAS asap.

(b)(6)(b)(7)(C)
(b)(6)(b)(7)(C) to Hauser, 2/1/06)

When Experiment 3 was recoded in February, 2006, Prof. Hauser's codes were used along with (b)(6)(b)(7)(C) coded the trials where (b)(6)(b)(7)(C) and Prof. Hauser disagreed, as tie breaker App. IV, P vs N, Item 3, [TAM_LastRun_0830.xls].

The significant finding from Experiment 3 that was reported in the *Nature* revision and discussed in the responses to a reviewer's comments, that tamarins successfully discriminated the Easy Ungrammatical items, disappeared. The p value changed from 0.029 to 0.40:

ok, so i ran the stats on experiment 3, and it failed. no discrimination on even the easy ones. here are the final stats. i will send to (b)(6)(b)(7)(C) and we will revise. it doesn't change the overall picture, but the expt 3 case was better with the successes on easy. that said, now we have accurate coding, and in the end, the right picture. thank goodness they succeeded on expt 1.

Thanks to all for helping out with the coding. now we have to figure out why they have crapped out on us.

marc

Experiment 1, (b)(6)(b)(7)(C), Final Statistics, February 15, 2006

Predictive: $z = 3.19$, $p = 0.001$

13 wins, 0 losses, 3 ties

Non-predictive: $z = 0.40$, $p = 0.69$

6 wins, 6 losses, 10 ties

Experiment 2, (b)(6)(b)(7)(C) Final Statistics, February 15, 2006

Predictive: $z = 0.39$, $p = .695$, 9 wins, 11 losses, 2 tiesNon-predictive: $z = 1.77$, $p = .08$, 7 wins, 15 losses, 0 ties

Experiment 3, (b)(6)(b)(7)(C): Final Statistics, February 20, 2006

Grammatical vs Ungrammatical: $z = 1.11$, $p = .27$, 8 wins, 4 losses, 7 ties**Easy Grammatical vs Easy Ungrammatical: $z = .85$, $p = .40$, 8 wins, 5 losses, 6 ties**Hard Grammatical vs Hard Ungrammatical: $z = .54$, $p = .60$, 5 wins, 5 losses, 9 ties

(Prof. Hauser to (b)(6)(b)(7)(C), 2/20/06, emphasis added)

Having found that the coding process was misrepresented, the further question for the Committee regarding Experiment 3 is whether Prof. Hauser intentionally miscoded the results of Experiment 3 in order to report a significant result on the "easy grammatical vs. easy ungrammatical" test.

When (b) (6) (b) (7)(C) was seeking to convince Prof. Hauser to permit the recoding of Experiment 3, he described how susceptible it was to miscoding:

Experiment 3 worries me because it hinges on the result with easy items, but there are only 2 ungrammatical and 2 grammatical items. A lot hinges on a little data, and just a few codes can make a big difference. This is especially important for Experiment 3, where the significant effect was small and the non-significant effect was almost significant. We cannot really trust these statistics until we do a full and proper coding. At the very least we owe it to ourselves to figure out exactly who coded what, since the same sort of confusion that occurred in Experiment 1 appears to have occurred in Experiment 3. (b) (6) (b) (7)(C) to Hauser, 2/5/06)

The Committee notes that Prof. Hauser repeatedly told (b) (6) (b) (7)(C) that they need not code Experiment 3; all they had to do was get him the video clips. (b) (6) (b) (7)(C) did, as he mentioned in the email exchange that was discussed above:

when I left the other dvd was burning in the g5 machine next to the door of the workshop. I imagine it is still in the drive there. Look at the file names for the particular files as they are prefixed with EUG, EG, HUG, HG, etc. ... (b) (6) (b) (7)(C) to Hauser, 9/3/05)

Prof. Hauser faulted (b) (6) (b) (7)(C) for this: "the files shouldn't have been marked as EUG, EG, etc. as this associates the files with the conditions, and prevents blind coding." Hauser Response, 5/12/09, p. 57 (emphasis added). However, (b) (6) (b) (7)(C) was merely complying with Prof. Hauser's request:

where are the other files for (b) (6) (b) (7)(C) will it be clear what trials are hard/easy...(Hauser to (b) (6) (b) (7)(C), 9/3/05).

There are no computer files or coding sheets to show that Prof. Hauser created a separate set of video files without indications of conditions, or otherwise took steps to code Experiment 3 blind. When Experiment 3 was recoded, Prof. Hauser's codes were used as one set, and (b) (6) (b) (7)(C) carried out a blind coding. (b) (6) (b) (7)(C) "tie broke" the cases where the two codings differed. In sixteen cases, the tie break favored (b) (6) (b) (7)(C) code, and in six cases it favored Prof. Hauser. In almost every case, the tie breaks where (b) (6) (b) (7)(C) code prevailed moved the result away from significance.

Examination of the video clips themselves shows a number of codes on "easy" trials where results were reported by Prof. Hauser that are clearly in violation of coding criteria, such as: clip 3, scored as a "No" when the subject drops out of sight during

playback (Grammatical Easy); clip 68, scored as a "No" though the monkey was moving constantly, and looked, so if it was not a "Bad" it was a "Yes" (Grammatical Easy); clip 31, scored as a "Bad" when it should have been a "Yes" (Grammatical Easy). All of these controverted codings were scored in the direction of significance by Prof. Hauser.

The Committee finds on the basis of the foregoing evidence that Prof. Hauser intentionally falsified the reported results of Experiment 3 in order to support the hypothesis he sought to present in the *Nature* manuscript.

§93.313 (f) (3) This research received PHS support. NIH/NIDCD sponsored award number_1 R01 DC005863-01A1. Funds from this award provided support for the tamarin colony.

§93.313 (f) (4) The Committee is not aware of any publications that need correction or retraction as a consequence of this misconduct.

§93.313 (f) (5) The Committee finds that Prof. Hauser is the person responsible for the misconduct.

(b) (6) (b) (7)(C)

(b) (6) (b) (7)(C)

The Committee is unaware of any pending applications by Prof. Hauser for federal support.

(b) (6) (b) (7)(C)

"Rhesus Monkeys Correctly Read the Goal-Relevant Gestures of a Human Agent," *Proceedings of the Royal Society B* (PRSB), 2007

Prof. Hauser directed research that members of his lab, graduate student (b) (6) (b) (7)(C) carried out on the island of Cayo Santiago, Puerto Rico, home to a colony of rhesus monkeys. The Committee's findings, set forth in detail below, are that Prof. Hauser did engage in research misconduct by recklessly or intentionally falsifying the research record, misrepresenting the results for one experimental condition, and misrepresenting the experimental procedures by reporting that all trials were videotaped.

The goal of the research was to determine whether rhesus monkeys were able to understand communicative gestures performed by a human. Preliminary experiments were run in 2005, and then trials were run in 2006 and 2007 that were reported in *PRSB*: "Rhesus monkeys correctly read the goal-relevant gestures of a human agent." The experimenter shows a subject monkey two boxes, and places the boxes several feet apart. Then the experimenter gestures towards one box and walks away. Conditions are varied by using different gestures, and by incorporating a display of a slice of apple in the presentation of the boxes in certain conditions.

(b) (6), (b) (7)(C) Prof. Hauser (b) (6), (b) (7)(C) observed and videotaped trials being run, and eventually decided against participating. While (b) (6), (b) (7)(C) expressed concerns about the manner in which the trials were conducted, the Committee does not rely on (b) (6), (b) (7)(C) observations or opinions in making its findings.

As a preliminary matter, the Committee notes that Prof. Hauser questions the credibility of (b) (6), (b) (7)(C). He cites, among other things, her denial of the particulars of a meeting about this project, and the corroboration of his account by (b) (6), (b) (7)(C). The source of the dispute appears to be that the witnesses are referring to different meetings, or they are conflating two meetings.

The issue arose when Prof. Hauser asked (b) (6), (b) (7)(C) to provide the CPC a letter regarding PRSB, and (b) (6), (b) (7)(C) replied that she did not recall seeing "the videos that (b) (6), (b) (7)(C) ran solo...I think that I had decided not to be on the paper before we actually found a time to meet to go over them together." Response 2/8/08, p.156. It appears that the meeting she is referring to is her meeting with Prof. Hauser on September 1, 2006; she acknowledges that she received a dvd of clips at that meeting, but says that she had already decided against being on the paper, so they did not review the clips together. The Committee has obtained a copy of the dvd from (b) (6), (b) (7)(C), and it matches files found on (b) (6), (b) (7)(C) computer. The dvd's content is consistent with what (b) (6), (b) (7)(C) said he would prepare: "Once I have all the video digitized and cleaned up I will send away with an excel sheet of my codes and the side I pointed to or looked at." (b) (6), (b) (7)(C) to Hauser, (b) (6), (b) (7)(C), 8/11/06). Immediately after that meeting, (b) (6), (b) (7)(C) is no longer copied on email by Prof. Hauser, (b) (6), (b) (7)(C). E-mail from August 29, 2006 through September 1, 2006 corroborates the contention of (b) (6), (b) (7)(C) that (b) (6), (b) (7)(C) did not attend that meeting.

Although there is no definitive record, it appears that there was a meeting several months earlier that (b) (6), (b) (7)(C) did not recall, which was the one described by Prof. Hauser:

After (b) (6), (b) (7)(C) ran the three follow-up conditions, I invited (b) (6), (b) (7)(C) to Harvard to a) view video footage of two additional experiments conducted by (b) (6), (b) (7)(C) using a video camera placed upon a tripod; (b) view the video footage of the three additional experiments conducted by (b) (6), (b) (7)(C) and filmed by (b) (6), (b) (7)(C); and c) join in as a co-author on the work, contributing both to the design of the subsequent conditions, as well as in the running and analyses. Though in an email correspondence (b) (6), (b) (7)(C) denied coming to Harvard to view these tapes with us, there is no question in my mind that on May 10, 2006, (b) (6), (b) (7)(C) did, in fact, come up to my office in William James Hall, sat down and viewed these tapes with myself and (b) (6), (b) (7)(C), and after some discussion, some with (b) (6), (b) (7)(C) as well, fully agreed that the experiments had succeeded. To confirm my recollection of this, I have included statements at the end of this section from both (b) (6), (b) (7)(C) about this, the videotapes that we reviewed at that time (see DVD marked by (b) (6), (b) (7)(C)) and email correspondence from the May 2006 time period. (b) (6), (b) (7)(C) later removed her name from the paper. Response 2/8/08, p.32, emphasis added.

(b) (6) (b) (7)(C) provided letters to support Prof. Hauser's position, Response 2/8/08, §2.2.10, p.47-48. While both Prof. Hauser and (b) (6) (b) (7)(C) claim the meeting was on May 10, that could not be correct because (b) (6) (b) (7)(C) had not returned from (b) (6) (b) (7)(C). On May 11, 2006, he emailed Prof. Hauser about more results on this very project: "They went 15 for 20 today..." The Committee has reviewed the dvd marked "For Marc (b) (6) (b) (7)(C)" and it contains 40 clips, all of one experiment, dated as having been created May 4, 2006 (not the filming date, but the clip creation or clip burning date). Even though the meeting could not have happened when Prof. Hauser and (b) (6) (b) (7)(C) say it did, and even though the dvd contained less extensive material than Prof. Hauser describes, it is reasonable to conclude that a meeting did occur in May, where (b) (6) (b) (7)(C) did watch some clips, and did not withdraw from the project. However, given the state of the record, where there were two different meetings each with an associated dvd of communicative gesture clips, the Committee does not find that the disagreement undermines the credibility of (b) (6) (b) (7)(C), or of any of the other participants.

There are two clear discrepancies between the research record and the *PRSB* article. One of the conditions involved pointing, without food. The paper reports that 31 monkeys out of 40 approached the target box in this condition. The spreadsheet that (b) (6) (b) (7)(C) prepared and copied on a disk with video clips for (b) (6) (b) (7)(C) when she visited September 1, 2006 only showed 27 subjects, not 31. Furthermore, only thirty video clips exist for this condition, corresponding to the first thirty trials coded on the spreadsheet. The paper reports that all trials were videotaped, and Prof. Hauser responded to a reviewer's concerns about experimental methodology by mentioning that all trials were videotaped *PRSB*. 2007, p.1914; Response_PRS_Reviews-2-15-07, Response_2ndReview_PRS³.

³ That having been said, the way the reliability was done on aborted trials in the new manuscript leaves much to be desired. One of the authors "selected" 20 aborted and 20 successful trials - we are not told how he selected these. And then he cut them so that the actual successful approach was not visible - but again we are not told how this happened, and that is of course critical. The trials were selected randomly from our data base. The randomization was done among the aborted and successful trials, respectively. Thus, (b) (6) had records of all trials run, classified according to whether they were aborted or successful. From these, he randomly selected 20 of each, across the full range of conditions....

This issue is critical for two reasons. First, keeping track of everything carefully and doing reliability carefully are just good scientific practice. Second, the authors compare their results to those of others studies with captive animals, but in those other studies there are exactly 0 aborted trials. I am not saying it should be done like this, but if aborted trials were counted as incorrect - on the assumption that if the subject knew where the food was he would go get it - then these results are probably not any different than those from other labs. My guess is that subjects failed to approach and engage in the task for several reasons, one of which was a lack of understanding of the communicative cue. The basic point is that we cannot know why subjects did not approach - at least not from the current study - and so the comparison to other studies is difficult.

We appreciate the reviewer's emphasis on tracking subjects, and we hope that the clarifications above are satisfactory. In terms of the contrast between field and captivity, we respectfully disagree with this reviewer's claim that in studies of captive animals, there are never aborted trials. MH has worked with captive primates for over twenty years, and all experiments have aborted trials, including failed presentations by the experimenter, distraction by the

Prof. Hauser's defense is that (b) (6), (b) (7)(C) recoded and the results changed from 27 to 31 successes:

While (b) (6), (b) (7)(C) was working on Cayo Santiago, he reported in an email that 27 out of 40 trials were successful in a pointing condition, a statistically significant result (binomial probability, $p < .02$); this is the email that the Committee refers to in their discussion of the allegation. After returning to Boston and recoding the videos of the trials, (b) (6), (b) (7)(C) final analysis included a result of 31 successes out of 40 trials, which is also a statistically significant result ($p < 0.0003$). Thus, the 27 successes out of 40 trials reflected the preliminary count taken 'online' in the field, whereas the 31 successes out of 40 trials reflected the final analysis after (b) (6), (b) (7)(C) looked over the videos and written records of the experiment. Let me further note that the counts can change during the final analysis based on closer examination of the videotapes – for example, it is common for subjects to be aborted during close inspection of video clips, based on the criteria listed in the methods section of the paper.

(b) (6), (b) (7)(C) and I recognize (see transcripts from interviews with (b) (6), (b) (7)(C)) that the data were poorly archived and so we cannot detail exactly which trials were aborted and why, or how they match up with the subset of video trials that the Committee refers to. These 30 video trials may reflect some of the data reported in the paper or they may reflect pilot data that (b) (6), (b) (7)(C) collected before beginning the experiment. (b) (6), (b) (7)(C) does not remember when each video clip was recorded and at what stage of the project, and so we cannot specify which data set these 30 video clips came from. Nonetheless, all trials were conducted using the same methods. Response, 5/12/09, p. 21, emphasis added.

There are no records of recoding by (b) (6), (b) (7)(C), and there are no email references to (b) (6), (b) (7)(C)

subject, equipment failure, and so; that some authors don't report this information is a different issue. More importantly, we did not count aborted trials as "incorrect." They were simply not counted in the data set. Thus, when we report the percentage of subjects selecting the correct box, this value is derived from the total number of subjects tested who watched the presentation and approached one box, picking either the correct or incorrect box in terms of the target action.

The reviewer does suggest that if we had counted aborted trials as "incorrect," on the assumption that failures to select either container demonstrate a failure to read the communicative gesture, then perhaps our data would look similar to those obtained with captive animals. Although this point is well taken, the majority of aborted trials are not cases where subjects fail to approach. Rather, most occurred because the subject was visibly inattentive, the trial was interrupted by another monkey, or the subject walked away before the end of the presentation...

At what point was the camera switched on?

The experimenter placed the tripod in position, once the subject was identified. The camera was immediately switched on. Thus, filming started before the experimenter started the presentation, and ran until the trial's completion.

carrying out any recoding. Prof. Hauser says that the result changed "after [REDACTED] looked over the videos and written records of the experiment." This is not credible, since the only existing records are the spreadsheet reflecting a total of 27 successes, and video clips corresponding to trials 1 through 30. It would be highly unusual to recode a set of video clips after a quarter of them had disappeared, and not communicate with the rest of the research team when the loss was noticed.

While Prof. Hauser suggests that "counts can change during the final analysis based on closer examination of the videotapes," he has argued the opposite case to the CPC:

The lack of ambiguity in the approach measure that we used for the Proceedings 2007 paper and the Science 2007 paper can be seen clearly in the video clips of these studies: unlike coding of looking- time or orienting responses that require a trained experimenter to code, it is trivial for an untrained observer to recognize which of the two boxes a subject touches first because the boxes are separated by some 2 meters and the subject cannot contact both boxes at the same time.
Response, 2/8/08, p. 29.

Reviewing the video clips in question at the time, while serving as a second coder, Prof. Hauser did not think that counts would change:

[REDACTED]

here are the codes for the trials you sent. very clear. i will be amazed if we have a single trial that differs. do you have videos of you doing the different conditions from the monkey's perspective? if not, please create these even in the lab so that we can have these on file and post them when the paper goes off.

marc

Condition	FileName	Box Selected
BASIC LOOK	3.MOV	L
BASIC LOOK	7.MOV	L
BASIC LOOK	13.MOV	R
BASIC LOOK	16.MOV	L
BASIC LOOK	24.MOV	R
COMM GEST	2.MOV	R
COMM GEST	12.MOV	R
COMM GEST	14.MOV	R
COMM GEST	29.MOV	L
COMM GEST	36.MOV	R
POINTING	2.MOV	R
POINTING	5.MOV	L
POINTING	19.MOV	L
POINTING	25.MOV	L
POINTING	26.MOV	L
STAND BOX	6.MOV	R

CONFIDENTIAL

Report of Investigating Committee
8 January 2010

STAND BOX	10.MOV	R
STAND BOX	31.MOV	L
STAND BOX	36.MOV	L
STAND BOX	42.MOV	L

(Hauser to (b) (6), (b) (7)(C) 9/3/06).

The Committee has examined the codes, and they match (b) (6), (b) (7)(C).

Finally, (b) (6), (b) (7)(C) performed further analysis on the data set; he does not recode, and the number of successes for the "pointing without food" condition remains 27:

On Sep 11, 2006, at 2:44 PM (b) (6), (b) (7)(C) wrote:

Marc and (b) (6), (b) (7)(C)

Comparing the with and without food communicative gesture conditions yields no evidence of learning. The communicative gesture condition with food has a binomial of .0011 (30/40). Subtracting the guys tested twice in the without food condition gives us a binomial of .0176 (24/34). The chi square comparing the two conditions yielded a p value of .670.

The same comparison between the with and without food pointing conditions also shows no evidence of learning. The with food pointing condition has a binomial of .00034 (31/40). Subtracting the guys tested twice in the without food condition gives us a binomial of .0147 (22/31). The chi square comparing the conditions gives us a p value of .530.

So, unfortunately for the Rhesus, they're not learning a damn thing! So we'll move on to writing up the methods and results.

(b) (6), (b) (7)(C)

From: (b) (6), (b) (7)(C)
Date: September 15, 2006 12:21:22 PM EDT
To: (b) (6), (b) (7)(C)
Subject: Re: quick statistical run down

Dear (b) (6), (b) (7)(C),

In the with/without food communicative gesture conditions there were six repeat tests and all six succeeded in both conditions. In the with/without food pointing conditions there were nine repeat tests three succeeded in both, three succeeded in food condition but failed without food condition, one failed both conditions, and two failed the with food condition and succeeded the without food

condition.

Reviewing these two emails, the starting point for tallying subjects who succeed in the "pointing without food condition" is 22, the number who were not tested twice, mentioned in the first email. In the second email, 3 subjects succeeded in both with and without food trials (22 + 3); and two failed with food but succeeded without (22 + 3 + 2 = 27). Since [REDACTED] mentions "writing up the methods and results" in the first email, it is highly unlikely that any recoding was done later.

Based on the record set out above, the Committee finds that Prof. Hauser recklessly or intentionally falsified the data point for the "pointing without food condition" by reporting it as 31 when it was 27. While Prof. Hauser is correct when he notes that both 27 successes and 31 successes yield a statistically significant result, it is also true that the misrepresentation strengthens the result, creating a more perfect match with the "pointing gesture with food" that also had 31 successes, and a sharper contrast with the "basic gaze" condition, see *PRSB* 2007, fig. 2.

The "pointing without food" condition is not the only one where the video record is missing. There are no video records at all for the 'communicative gesture with food' condition, and there are very few recordings of aborted trials. The absence of video records is a serious matter because Prof. Hauser met the criticism of reviewers by representing that all trials were videotaped:

While the idea of this study is a good one, the execution is very poor. It does not follow accepted experimental procedures. The problem is that there is a single experimenter deciding on the spot whether or not a trial should be run based on the potential subject's behaviour. Amazingly, the authors report that "approximately 60% of trials were aborted". There is no assessment of inter-observer reliability on this judgement - and indeed, how could their be, as they only know "approximately" how many trials were aborted. Assessment of inter-observer reliability is mandatory as, obviously, the aborted trials could easily be ones in which the subjects were likely to fail. In the laboratory studies using this paradigm, there are typically 0 trials that are aborted. There is also no assessment of inter-observer reliability on any of the rest of the experiment, again an unacceptable experimental procedure. ... While some field studies cannot perform inter-observer reliability for practical reasons, the current experiments could easily be conducted by two observers or videotaped. (PRS First Review, 2/7/07, emphasis added.)

>>>>>

...

Specific response to Referees' Criticisms:

First, both referees are concerned that we were consciously or unconsciously biased in our coding of successful versus aborted trials. This is a valid concern, and here is how we have addressed this problem. [REDACTED] who ran

the experiments, filmed each trial from a remote video camera. (b)(6)(b)(7)(C) took 20 trials that were considered successful trials (i.e., the animal was allowed to pick one of the two boxes and we included the data in the analyses) and 20 trials that were counted as aborted trials (i.e., not included in the analyses), and handed these to (b)(6)(b)(7)(C) who scored them blind to the abort/success label. That is, (b)(6)(b)(7)(C) simply coded based on our criteria whether a trial should be counted as an abort or as a successful trial. There was 100% agreement. Thus, when (b)(6)(b)(7)(C) was deciding whether to abort a trial, online, in the field, he was not unconsciously biased in this decision. If he had been biased, then (b)(6)(b)(7)(C) blind coding would not have corresponded to (b)(6)(b)(7)(C) rating. ... (Hauser to PRSB Editor, 2/16/07, emphasis added)

Since Prof. Hauser not only reported in the article that all trials were videotaped, but made that representation in responding to referees' comments, it was reckless for him to have failed to confirm that the research team did videotape all trials, or to have failed to secure them, if they did exist at some time.

In the same correspondence to the editor, Prof. Hauser also addressed a concern about a related bias problem:

"c. Aren't the author's selecting the best subjects, i.e. those most likely to respond to human cues, as they took part in the experiment, probably having the right temperament, greatest curiosity and least neophobia."

As noted above, our procedure for finding test subjects was not based on prior performance, but rather, on finding lone individuals. Any individual who was alone, paid attention to the presentation, and approached the box was included. It is, of course, possible, that only curious and minimally neophobic animals would do this. That said, we sampled a relatively large number of individuals over the course of the several conditions, and even if we only picked curious/non-neophobic animals, the goal here was to show that some rhesus can do this. We don't think we have a biased sample. (Hauser to PRSB Editor, 2/16/07, emphasis added)

The reassurance about subject selection that is highlighted appears to be contradicted by this email correspondence between Prof. Hauser and (b)(6)(b)(7)(C). The exchange, following up on the trials run by (b)(6)(b)(7)(C) with (b)(6)(b)(7)(C) observing, suggests that there was an attempt to identify "the best subjects:"

From: Marc Hauser <mdh@wjh.harvard.edu>
Date: December 10, 2005 3:32:02 PM EST
To: (b)(6)(b)(7)(C)
Subject: Re: gaze

well, that is a drag. and what a lesson here. in the past, when we have done stuff with 40 subjects, we would pop champagne. so, glad we spent the extra time. i hope that we can still pull somethign out of this. what i would like us to try is to get back the dominant guys and rerun with them. we shouldn't do this until we finish th elemon study, but i want to see if we can get a robust effect with repeating guys who have been run and are dominant. this would

be of interest to many other studies.

yes, we got a bunch, but melting today and then icing, so it sucks.

marc

On Dec 10, 2005, at 1:07 PM (b) (6), (b) (7)(C) wrote:

Marc,

So they went 5 for 10 and (b) (6), (b) (7)(C) felt we should call it and I agreed. We didn't test any monkey we tested in the earlier trials. **My feeling is that as we test monkeys that are lower in rank than the first few trials of guys we tested, we are testing guys that are more ancy.** (b) (6), (b) (7)(C) and I thought this might be a possibility when we were trying to figure out why some clearly get it and others seem not to. So today we had to settle for guys with whom I couldn't get that great eye lock that I was getting with nearly all in the first few conditions (but I did get it with some today).

It felt so random today that compared even with the first condition (b) (6), (b) (7)(C) and I ran, I think there must be some variable that would bring them from 16,17 and 14 of 20 all the way down to chance. I'm going to take a look at the video from today and last time and see if I can find some sort of rhyme or reason as to who chose correctly and who did not. I'll look specifically at the guys who B-line and see if I can find a pattern. (b) (6), (b) (7)(C) suggested that it might be worth it to look at not only who approaches, but who hesitates before they approach and how long. Also, I'll send you a video of a couple guys who do the B-line choice and a few who choose incorrectly so you can have an idea of what this looks like.

After I look at the video I'll email you and (b) (6), (b) (7)(C) and tell you if I see anything interesting. Does this sound ok? I read that Logan got almost 9 inches of snow which is a lot for December. January and February ought to be tons of fun! Talk to you soon.

(b) (6), (b) (7)(C)

(Hauser to (b) (6), (b) (7)(C) to Hauser, 12/10/05, emphasis added)

Yet another bias problem is illustrated by this email and Prof. Hauser's Response, in the section discussing the importance of replications of this study, where he says: (b) (6), (b) (7)(C) (b) (6), (b) (7)(C) collected 20 trials in the 'communicative gesture with food' condition and replicated the general pattern (14 out of 20 subjects), although this just misses the traditional level of statistical significance in this field ($p = .057$).” Response 5/12/09, p.22. (b) (6), (b) (7)(C) did not collect 20 trials with 14 successes; they collected 30 trials with 19 successes, which misses significance by a considerably wider margin. (b) (6), (b) (7)(C) explained that on the second day of her visit she was videotaping from very far away, addressing (b) (6), (b) (7)(C) concern from the day before that her proximity might be affecting results. (b) (6), (b) (7)(C) Interview, 11/11/08, p.25. On that day “they went 5 for 10;” and then (b) (6), (b) (7)(C) agreed with her suggestion to stop. Prof. Hauser unaccountably ignores the second day results, permitting him to claim that the

(b) (6), (b) (7)(C) trials help his case, when in fact they do not. (b) (6), (b) (7)(C) left the videotapes of the trials with (b) (6), (b) (7)(C) but they have been lost. Numerous pilot trials and trial repetitions took place in the course of this project. It is not possible to reconstruct a record of all trial runs in order to compare what trials were reported to the balance of trials that were not, to know whether sets of trials should have been included, but were not.

(b) (6), (b) (7)(C) took field notes for the trials, but they have been lost or discarded. Prof. Hauser has taken the position that field experiments need not be videotaped. That position must be founded on the assumption that appropriate field notes are systematically recorded for every trial, consistently capturing all relevant data, and those notes are carefully protected. No protocol-specific templates for field notes have been located. Prof. Hauser has admitted that (b) (6), (b) (7)(C) note taking and record retention practices were slipshod. Response 5/12/09, p.21.

The Committee considers these flaws in the research record in order to determine whether *PRSB* 2007 "need[s] correction or retraction," §93.313 (f) (4). Another argument Prof. Hauser has made in support of the validity of Cayo Santiago research carried out by (b) (6), (b) (7)(C) is that he conducted trials blind to condition, (b) (6), (b) (7)(C) was not given "our specific hypotheses as to the results we expected" in the action experiments, Response, 5/12/09/ p.29).

The following email exchange from the pilot phase of the project demonstrates not only that (b) (6), (b) (7)(C) was aware of possible selection criteria that would affect performance (gender and dominance) as well as the hypothesis. Further, the emails show Prof. Hauser flagging the parameters of statistical significance for (b) (6), (b) (7)(C)

From: Marc Hauser <mdh@wjh.harvard.edu>
Date: November 9, 2005 4:11:56 PM EST
To: (b) (6), (b) (7)(C)
Subject: Re: wednesday

(b) (6), (b) (7)(C)

let's run another 20 and 20 with the new method; rerunning animals you have run would be fine too. 16 out of 20 is significant and 11 out of 20 is not. but let's do it again with the more exaggerated effect. we want to be absolutely sure.
then we can move on to causality.
marc

On Nov 9, 2005, at 3:55 PM, (b) (6), (b) (7)(C) wrote:

Marc,

I just read your email about the clip. I forgot to check this morning. Odds are I can run 20 males and 20 females tomorrow being far more exaggerated with the whole thing. Sorry about not checking for your response. So should I try and get in the 40 trials tomorrow with the exaggerated looking or should I move on with causality? Either way I'll

make a clip now and send it.

Today, the last male succeeded which brings them to 16 successful trials out of 20, and the females succeeded (chose the apple box) in 11 of 20 trials. So the females weren't exactly terrified of me, but it seems like there might be some effect. Is 11 out of 20 vs. 16 out of 20 enough of a significant difference? And again some of the higher ranking females would b-line straight for the apple box suggesting that some of them understand what is happening.

(b) (6), (b) (7)(C)

(Hauser to (b) (6), (b) (7)(C) to Hauser, 11/9/05)

The Committee finds that, in the absence of a clear and complete research record, it was reckless for Prof. Hauser to rely on the results reported by the RA who performed the experiments after being educated about the hypotheses and the success rate that would achieve statistical significance. Success in subsequent related experiments may show that there was not gross fabrication on (b) (6), (b) (7)(C) part, but it does not fill the inexcusably large gaps in the research record that Prof. Hauser overlooked as he submitted and revised the *PRSB* manuscript for publication. One measure of the standards of the relevant research community is to consider whether, had the editor been informed in advance of the state of the research record of this project, it could have been published. In our view, it probably could not have been. Prof. Hauser has stated as much in a similar context: "Given the mandate by all journals that published work is open to scrutiny by peers, and given that I and my students...were well aware of this, we certainly knew full well that the absence of evidence would be disastrous for any claims we wished to make." Hauser Response 5/12/09, p. 14. The Committee finds that, under the provisions of §93.106, the failure to maintain records is evidence of research misconduct.

§93.313 (f) (3) This research received PHS support. Grant number CM-5-P40RR003640-13 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), provided funds to maintain the rhesus colony on Cayo Santiago. (b) (6), (b) (7)(C) work was supported by NIH National Research Service Award (NRSA) grant F31MH075298.

§93.313 (f) (4) The Committee finds that Prof. Hauser should notify the editors of *Proceedings of the Royal Society B* regarding "Rhesus monkeys correctly read the goal-relevant gestures of a human agent," to inform them that the results for one condition should be corrected, and that videotapes do not exist, and may never have been made, for trials reported in the article and for aborted trials so that the editors can decide whether correction is sufficient, or whether retraction is called for.

§93.313 (f) (5) The Committee finds that Prof. Hauser is the person responsible for the misconduct.

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

The Committee is unaware of any pending applications by Prof. Hauser for federal support.

(b) (6), (b) (7)(C) "The perception of rational, goal-directed action in nonhuman primates," *Science* 317, 1402 (2007)

Experiments with three different species were reported in J. Wood, D. Glynn, B. Phillips, and M. Hauser, "The perception of rational, goal-directed action in nonhuman primates," *Science* 317, 1402 (2007). One set of experiments was conducted with cotton-top tamarins in William James Hall; a second with rhesus monkeys on Cayo Santiago; a third with chimpanzees at the Tchimpounga Sanctuary in the Congo.

Cotton-top tamarin experiments in William James Hall

These experiments were performed by (b) (6), (b) (7)(C) and (b) (6), (b) (7)(C) in the Hauser Lab. (b) (6), (b) (7)(C) did not assist with the coding of the experiments, which was done by (b) (6), (b) (7)(C) in the lab. She was sent copies of the draft paper to review and provided only a few comments. No concerns have been raised about the data collection or reporting for these experiments and the Committee did not review this material.

Rhesus experiments on Cayo Santiago

(b) (6), (b) (7)(C) was the sole research assistant who ran the experiments on Cayo Santiago and collected all the rhesus data reported in the *Science* paper. In addition to (b) (6), (b) (7)(C) Prof. Hauser and (b) (6), (b) (7)(C) were involved with experimental design and data analysis. The protocol on Cayo Santiago involved presenting a monkey with two upside-down coconut shells (attached to wooden panels for ease of handling) and performing an action on one of the shells: in Experiment 1, (b) (6), (b) (7)(C) performed either an "intentional" action, a direct hand grasp of the shell, or an "accidental" action—a casual touching by "flopping" the back of the hand on the shell; in Experiment 2, (b) (6), (b) (7)(C) touched one of the shells with his elbow, either with hands empty or with one or both hands occupied (holding a piece of cloth).

Data from the rhesus experiments do not exist: there are no field notes, videotape recordings, records of abortions, or identifying information about subjects tested. (b) (6), (b) (7)(C) have confirmed in interviews that no test trials of rhesus subjects were videotaped (also see Prof. Hauser's statements, 8 February 2008, page 38 and 12 May 2009, page

24). Prof. Hauser explains that “most of the observations were hand written by [REDACTED] on a piece of paper, and then the daily results tallied and reported to [REDACTED] over email or by phone” and then the raw data were discarded.

The total absence of data presents special difficulties in the review of this research.

Identification and repeat testing of subjects

[REDACTED] reports, in an email to Prof. Hauser (27 July 2007) discussing general methods in use on Cayo Santiago, that one half of the rhesus subjects tested cannot be identified.

Also, while I think we should really put out an effort to collect ID data, the tattoos are almost impossible to see for probably 40-50% of the monkeys. Perhaps the hair has grown over, or the tattoos need to be redone, but **in general I would say we can only collect IDs successfully for 50% of the subjects.** [emphasis added] Also, there are cases when you can see some part of the tattoo and can make a guess about about [sic] identity of the subject, but to what extent do we want people making even informed guesses? So, we might want to think of some way to deal with this problem, perhaps by only recording IDs when we are absolutely sure.

However the *Science* Supporting Online Material (page 6) states that “All individuals are well habituated to human observers and **readily identifiable by natural markings along with chest and leg tattoos and ear notches.**” [emphasis added]

Prof Hauser, in his 12 May 2009 response (p. 26), asserts that this statement does not contradict [REDACTED] email, nor is it misleading to readers of the *Science* article, because “my students and I often rely on natural markings (e.g. size, facial characteristics, coloring patterns, group membership, distinguishing characteristics) in addition to remembered IDs (when visible) to identify the individuals” and “it is not difficult for a researcher to recognize whether he has already tested a subject in an experimental condition, given that each condition is often carried out within a single day and for a trained observer, the monkeys look very different from one another. This recognition of previously tested subjects can easily be done on the basis of the natural markings and remembered IDs.” However the Committee finds that the clear implication of “readily identifiable” is that subjects’ identities can be, and were, ascertained, not that a trained experimenter might be able to remember whether a specific monkey had already been tested, if all trials were carried out in a single day.

The absence of information about repeat testing is relevant not only within conditions but between conditions. The *Science* paper implies that each subject was tested only once, and saw only one of the four conditions (Supporting Online Material, page 7, and Wood *et al.* cover letter to *Science* responding to reviewers’ comments, 7 June 07, item 1a).

Rhesus were tested in a between-subjects design, receiving either one intentional trial (n = 20 subjects) or one accidental trial (n = 20 subjects);

experiment 1), or one hand-occupied trial (n = 32 subjects) or one hand-empty trial (n = 32 subjects; experiment 2). [Supporting Online Material, page 7]

This information [whether testing was within-subject or between-subject] was in the original manuscript, but was obviously not clearly described. We now clarify these methodological issues in the SOM, specifying that the tamarins were tested in a within-subjects design, receiving multiple trials for each action type; the rhesus were tested in a between-subjects design, receiving a **single trial of one action type**; and the chimpanzees were tested in a within-subjects design, receiving a single trial of each action type. [cover letter, emphasis added]

The total number of trials reported in *Science* was 104. In addition, the paper reports there were 39 aborted trials.⁴ This implies that 143 trials were presented, but since there are no experimental records and since many subjects could not be identified it is not known which subjects were tested or whether any subject was tested more than once. That subject identities could not be reliably ascertained, and were not recorded, calls into question the declaration that rhesus received “a single trial of one action type.”

To defend against this difficulty, in his 12 May response to the Investigating Committee Prof. Hauser maintains that retesting would not affect subjects’ performance.

Concerning retesting, it is possible that a few subjects may have been tested in more than one condition. Because we did not record ID information in this particular study we cannot determine whether some subjects were tested in more than one experimental condition, each of which occurred several days apart. However, even if we did retest individuals in different conditions, it is clear from several studies that we have carried out, and as previously noted to the CPC, that retesting doesn’t change performance relative to results from naïve subjects in these experimental situations....[T]here is no evidence that in this population of rhesus monkeys, using such methods, that retesting has any significant effect on performance. [page 27]

Yet this claim is directly contradicted by other statements four pages later in the same response regarding subjects habituating over time. (“[P]atterns of response can change, both because **animals can habituate to the overall procedure**, or can lose sensitivity to certain dimensions of the procedure.” Page 31, emphasis added.) Since animals become habituated only through retesting, this would appear to be clear evidence that retesting can have a significant effect on performance.

Prof. Hauser also asserts in his 12 May response that “in a wide number of field studies, researchers typically carry out experiments without precise information on individual or

⁴ This represents a 27% abort rate. In his interview with the Committee on 16 March 2009, [REDACTED] said that the abort rate for the “coconut studies” was more than twice as high as reported in the *Science* paper: “I would say that on average, about 60% were aborted, although again it depended on the study. We did one study where the food containers were upside-down coconut shells and that tended to yield slightly more aborts than studies where we would actually show food and then drop it behind an occluder.”

group identity” and cites two studies as examples. However one cited study was of vocal responses of naturally-occurring, isolated groups of monkeys, where the intention was explicitly to test the same group repeatedly, not to ensure that individuals were not retested; the second study was of naturally-occurring groups “on a stable home range that is defended against neighbouring conspecific groups” where group locations were determined via GPS and map to insure against testing the same group more than once.

Problems with stimulus presentation

After the paper had been submitted to *Science* [REDACTED] was asked to create “monkey’s eye view” video clips demonstrating the procedure used with the rhesus subjects for inclusion with the Supporting Online Material. On 13 June 2007 he emailed four clips to Prof. Hauser and to [REDACTED] [REDACTED] email response indicates there were problems with [REDACTED] presentations. The demonstration video included in the Supporting Online Material—produced after the data had been submitted for publication—reflects [REDACTED] coaching.

Subject: Re: action videos

From: [REDACTED]

Date: Wed, 13 Jun 2007 17:01:47 -0400 (EDT)

To: [REDACTED]

CC: [REDACTED]

Hey [REDACTED]

These look good. I think there are a few things that we want to change. First, you should **make sure that you walk directly down the middle. You probably are in real life, but on the video it looks like you are going to one side.** [emphasis added] The best way to do this might be to do a practice trial, then look on the video and mark a spot between the shells 10 m back that looks right on the video. Second, for the hand-empty elbow grasp, **your hand should be to the side of your body, rather than in front of your body.** [emphasis added] Third, make the flop look less intentional. It looked like you **intentionally hit the coconut with the back of your hand, rather than letting your hand causally [sic] fall on the coconut.** [emphasis added] Otherwise, looks great.

[REDACTED]

On Wed, 13 Jun 2007, [REDACTED] wrote:

- > So I took these videos this afternoon as I figure
- > there are going to be little issues that we'll want to
- > correct the first time around, so fire away.
- >
- > [REDACTED]

(b) (6), (b) (7)(C) had been the only experimenter to test subjects by performing these actions, and had conducted more than 100 trials (perhaps significantly more, since experiments were routinely pilot tested⁵). (b) (6), (b) (7)(C) comments indicate that the experimental actions that (b) (6), (b) (7)(C) actually used—as performed by him on the videos sent to Hauser and (b) (6), (b) (7)(C)—may have differed in important ways from those reported in the paper (See *Science* paper, page 1403; and Supporting Online Material, clip S2 and “Rhesus” section). Prof. Hauser acknowledged the importance of even “subtle differences” in his 8 February 2008 statement (page 30).

Given the subtle nature of these experiments, and given the fact that these experiments often require that the subjects notice subtle differences in the experimenter’s actions, **it is critical that they be performed correctly in all respects in order to generate interpretable data.** [emphasis added]

In his 12 May 2009 response, Prof. Hauser states that (b) (6), (b) (7)(C) was asked to redo the sample stimulus movies because details of the gestures were difficult to see within the movies.” In fact the gestures are quite clear, and no less clear than gestures ultimately included in the Supporting Online Material.

While conducting the experiments reported in the *Science* article, (b) (6), (b) (7)(C) had performed the same procedures he recorded on videotape perhaps hundreds of times. That the procedures differed in obvious ways from those reported in the article calls into question the interpretability of the data obtained. It is noteworthy that no such concern was raised by (b) (6), (b) (7)(C) and Prof. Hauser in response to the “monkey’s eye view” videos sent by (b) (6), (b) (7)(C), and the incident underscores the importance of keeping a video record of experiments. With no video records of the actual experiment to examine, (b) (6), (b) (7)(C) had no basis for instructing (b) (6), (b) (7)(C) on the fine points of his reenactment, since it was (b) (6), (b) (7)(C) not (b) (6), (b) (7)(C), who had performed over a hundred trials.

Potential experimenter bias

Prof. Hauser’s statement of 8 February 2008, discussing the *Science* experiment on Cayo (page 38), asserts that (b) (6), (b) (7)(C) was not informed of the specific hypotheses of the experiments before he conducted them. Thus, he tested the monkeys in these conditions “blind” to the hypothesized result.” Contrary to this assertion, there is clear evidence in the email record, starting from (b) (6), (b) (7)(C) first days on Cayo (he arrived on 11 July 2005), that he was not only informed of experimental hypotheses of projects on which he worked (and of the outcomes that would support those hypotheses), but often assisted in the development of those hypotheses. Several examples from the email record are included in Appendix Folder VI, items 1 and 2.

Prof. Hauser denies this interpretation in his 12 May 2009 response, stating that (b) (6), (b) (7)(C) was not aware of specific hypotheses because (b) (6), (b) (7)(C) had not provided him with the relevant literature.

⁵ (b) (6), (b) (7)(C) interview, 10 February 2009, pp 10-11.

We certainly did not claim that (b) (6) (b) (7)(C) was blind to the broader hypotheses we then held for all of our experiments (which would be impossible in any case as the actions are often transparent regardless of the relevant hypotheses), nor did we claim this in the paper. Rather, for a subset of the action experiments (b) (6) (b) (7)(C) provided (b) (6) (b) (7)(C) with experimental methods, but not with the relevant literature that motivated these studies (i.e. the previous studies with human infants on which our study were based; Gergely, Bekkering & Kiraly, 2002, *Nature* 415: 755), nor with our specific hypotheses as to the results we expected.

Whether or not he had been provided all the relevant literature, (b) (6) (b) (7)(C) a lone experimenter with little training, was solely responsible for forwarding data collected by himself from trials where he was aware of the "desired" outcome (in this case, that rhesus monkeys would "get it" when the action performed on an object was obviously intentional and/or goal-directed vs. accidental), in a study where trials were not being recorded, and where he alone was judging whether a trial should remain part of the record or be classified an abort. The Committee finds that the opportunities for both conscious and unconscious experimenter bias presented by this situation are undeniable, and that it was irresponsible and reckless of Prof. Hauser to allow such work to be represented as methodologically sound and to be submitted for publication.⁶

Contradictory results not acknowledged or reported

After the paper had been submitted to *Science*, but before a decision had been made to publish, (b) (6) (b) (7)(C) ran additional trials with the inverted coconut shells as containers. Although Prof. Hauser states that the results contradicted the interpretations presented in the *Science* paper, no attempt was made to modify or update the reported findings.

Subject: update
Date: Mon, 11 Jun 2007 15:01:16 -0400
From: (b) (6) (b) (7)(C)
Reply-To: (b) (6) (b) (7)(C)
To: (b) (6) (b) (7)(C)

Marc,

The elbow condition I ran yesterday was with cococuts/hands empty/no looking at all, which I thought was what we had all discussed, but (b) (6) (b) (7)(C)

⁶ Indeed, Prof. Hauser acknowledges the likelihood of experimenter bias in his 12 May 2009 response at page 29: "Similarly, in studies of human infants, such as those by my colleagues (b) (6) (b) (7)(C), the experimenter testing the infant knows in advance the predicted results and relevant theories. Although they take great precautions against biases, including rehearsing the presentations so that they are identical across experimenters, it is possible for unconscious cueing to intervene." [emphasis added]

CONFIDENTIAL

Report of Investigating Committee
8 January 2010

said he thought it was supposed to be with my hands full.[...]

* * * *

From: Marc Hauser <mdh@wjh.harvard.edu>

Date: June 11, 2007 3:50:45 PM EDT

To: (b) (6) (b) (7)(C)

Subject: Re: update

Reply-To: mdh@wjh.harvard.edu

(b) (6) (b) (7)(C)

thanks for the update.

i think i am still puzzled on the first one. you placed 2 coconuts upside down, never made eye contact, and just indicated with elbow but with both hands free, and they were 15 out of 20? if that is the case, then i am totally confused as this goes against everything we have done. that is, they shouldn't have been successful on elbow unless both hands holding occluder. with both hands free, which wasn't any of our conditions (the contrast is with one hand holding occluder/cloth and one hand free), they should have failed. [emphasis added] also, without making eye contact, they should have failed too, at least based on gergely interpretation. so walk me through this one again.[...]

* * * *

From: (b) (6) (b) (7)(C)

Date: Mon, 11 Jun 2007 17:54:49 -0400 (EDT)

To: mdh@wjh.harvard.edu, (b) (6) (b) (7)(C)

Subject: Re: update

hey guys,

yeah, it is very weird that they are seeing the hands-empty action as goal-directed (even without eye contact!), given that we have two previous robust failures (7/16 subjects and 8/16 subjects [sic; the *Science* paper reports a total of 16/32 subjects on this condition, not 15/32]) using this same hand-empty elbow action with coconuts. i am worried that they are becoming increasingly sensitive to seeing our actions as goal-directed, perhaps because of the huge number of forced-choice action experiments we have run over the past few years. and many of these experiments have included baiting, which would only enhance this. it will be interesting to see what happens with the hand-empty bucket condition tomorrow; if they fail, then great, we've still got a method that works with these guys. but if they succeed, then i think we are dealing with increased sensitivity to goal-directed actions. [emphasis added] which, btw, is interesting from the broader mirror neuron perspective (that is, it's completely inconsistent with their account), but does cause us some problems. a similar thing

happened with the tamarins as well, in a different way (they stopped responding altogether, when they realized they never got food on action trials). so in the end we might have to get creative and think up some new methods.

what do you guys think? (b) (6), at one point on the phone you mentioned that they seemed to be going for everything, [emphasis added] which would support this idea. also, does it seem like they are being fed regularly?

(b) (6) (C)

(b)
(7)
(c)

* * * *

On 6/12/07 8:09 PM (b) (6), (b) (7)(C) wrote:
Guys,

They were 17 for 20 going for the elbow in the bucket condition. So this seems like a fairly straightforward result in that **they are just going up to anything we act on.** [emphasis added] Marc, I think it makes sense to move on to the apple rolling experiment.[...]

Ultimately, Prof. Hauser decided simply to ignore the contradictory results:

From: Marc Hauser <mdh@wjh.harvard.edu>
Date: June 12, 2007 8:14:26 PM EDT
To: (b) (6), (b) (7)(C)
Cc: (b) (6), (b) (7)(C)
Subject: Re: update

well, very interesting. yes, **let's back off this and turn to our new methods....**
[emphasis added]

In his 12 May 2009 statement, Prof. Hauser explains that this effect was not unexpected and therefore not worthy of remark. "[P]atterns of response can change, both because animals can habituate to the overall procedure, or can lose sensitivity to certain dimensions of the procedure.... Regardless of condition, a person is always in some way in contact with one container as opposed to the other. A likely outcome, therefore, especially after months of testing, is that subjects simply use raw association (the container that is paired with human proximity) to guide their approach behavior. Alternatively, because of the lack of a food reward for many conditions, subjects approach at random." [page 31]

Once again, this clearly implies that subjects' responses change over time as a result of retesting—and yet it contradicts Prof. Hauser's earlier assertion that retesting does not affect performance.

Retesting doesn't change performance relative to results from naïve subjects in these experimental situations....[T]here is no evidence that in this population of rhesus monkeys, using such methods, that retesting has any significant effect on performance. [12 May 2009 response, page 27]

The Committee finds that Prof. Hauser was aware of the potential effects of retesting, yet chose to acknowledge those effects only when the data did not support the experimental hypotheses.

Chimpanzee experiments at Tchimpounga Sanctuary

The Supporting Online Material for the *Science* paper indicates that 40% of the chimpanzee trials were videotaped.⁷ However, none of the fifteen subjects run by Hauser were videotaped, and only five of the ten subjects run by (b) (6), (b) (7)(C) were videotaped. Prof. Hauser, in his 28 April 2008 response, describes how the 40% figure could be deemed accurate:

The inquiry report notes that the *Science* 2007 article reports that “40 percent of the 25 trials were videotaped,” and then asserts that email and computer records indicate that only 20 percent of the trials were videotaped (IR 8). This finding is simply wrong. Here is the correct breakdown:

- **15 subjects, not videotaped (included in final analyses); Experimenter: M. Hauser**
- **5 subjects, not videotaped (included in final analyses); Experimenter: (b) (6), (b) (7)(C)**
- **5 subjects, videotaped (included in final analyses); Experimenter: (b) (6), (b) (7)(C)**
- **5 subjects, videotaped (excluded from final analyses due to side bias, n=4, and failure to participate, n=1); Experimenter: (b) (6), (b) (7)(C)**
- **3 subjects, videotaped (excluded from final analyses due to a complete lack of interest in the task, i.e. the session was never finished, n=2, or because they had been tested previously, n=1); Experimenter: (b) (6), (b) (7)(C) [note: these subjects are not mentioned in the article or in the Supporting Online Material]**

Thus, we tested 33 chimpanzees, and videotaped 13 (39.4%) of those subjects. This is equal to the 40% of trials that we reported to have videotaped.

The Committee finds that the statement about recording “a subset of the trials (40%)” clearly implies that 40% of the trials reported in the article were videotaped. The Supporting Online Material states that twenty-five chimpanzees participated. Only five subjects of those twenty-five (i.e., 20%) were videotaped. While this distortion may not itself affect the actual data reported, it does falsely

⁷ “A subset of the trials (40%) were recorded onto video and scored by two coders; inter-observer reliability was 100%.” Wood et al., 2007, *Science*, Supplementary Materials, p. 9-10.

represent the inter-observer reliability of the coding of those data.

Conclusion

There is a disturbing pattern of misrepresentation of results and shading of truth in the reporting of this research. On the one hand, none of the experiments on Cayo Santiago were recorded and no field notes exist, so that questioned findings cannot be reviewed or data re-analyzed. On the other hand, the total absence of raw data, reliance upon summary reports from a single unsupervised research assistant with whom Prof. Hauser shared experimental hypotheses and estimates of success rates needed for statistical significance, and unwillingness to acknowledge or pursue clear evidence challenging the integrity of the reported results before the paper was published, reflects a reckless disregard for basic scientific standards. Perhaps more troubling, responses to questions about apparent discrepancies or other problems with the report of the research have themselves been inconsistent or misleading. The Committee finds that Prof. Hauser's behavior with regard to the research reported in the *Science* article constitutes research misconduct under §93.103 and §93.106; in the absence of any research data, the Committee cannot state whether the misconduct was fabrication, falsification, or both. Specifically, Prof. Hauser breached his obligation to ensure that the rhesus experiment results reported by (b) (6), (b) (7)(C) had been recorded in field notes prepared at the time of the trials, and that those notes and spreadsheets based on those notes were preserved. In the absence of these records, and in the face of the contradicting results showing that "they are just going up to anything we act on," the Committee finds that Prof. Hauser was reckless in permitting the publication of the *Science* article.

§93.313 (f) (3) This research received PHS support. Grant number CM-5-P40RR003640-13 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), provided funds to maintain the rhesus colony on Cayo Santiago. (b) (6), (b) (7)(C) work was supported by NIH National Research Service Award (NRSA) grant F31MH075298.

(b) (6), (b) (7)(C)

§93.313 (f) (4) The Committee finds that Prof. Hauser should notify the editors of *Science* about the lack of documentation for the reported experiments on Cayo Santiago so that the editors can decide whether retraction is called for, or whether some other measures should be taken.

§93.313 (f) (5) The Committee finds that Prof. Hauser is the person responsible for the misconduct.

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

The Committee is unaware of any pending applications by Prof. Hauser for federal support.

(b) (6), (b) (7)(C) AXA

Introduction

AXA was a language playback experiment carried out with rhesus monkeys on Cayo Santiago. (b) (6), (b) (7)(C) assisted respondent with the design of the experiment, using the methodology of habituation followed by exposure to stimuli based on rhesus calls, that either conform to or violate the patterns of the stimuli strings used to habituate. The responses of subjects were recorded on videotape. Hauser Lab protocol called for the creation of digitized video clips in order to facilitate coding the trials blind to condition. (Letter of (b) (6), (b) (7)(C), p.4, Hauser submission to Committee, 12/10/08 at tab 7). No digitized video clips were found for this project.⁸ Digital video was found in the form of iMovies of each subject on the Maxtor 500 gigabyte external hard drive used by (b) (6), (b) (7)(C). While it would be possible to code blind from the iMovies, it would be appreciably more difficult, and contrary to laboratory protocol. The Committee finds that the absence of the video clips is some evidence that coding was not carried out blind to condition.

It is alleged that Prof. Hauser falsified research records by changing the coding of trials in the AXA experiment, intentionally reporting the responses of subjects in a manner that would yield a statistically significant result. Prof. Hauser submitted a follow-up memorandum after his interview in which he acknowledged that "there was a change from non-significance to significance, but the non-significant and original result was based on a completely different (and inappropriate) method of coding. The change in coding was completely justified based on the design of the experiment." The Committee finds that the change from non-significance to significance cannot be accounted for by the change in coding method, and that, as set forth in detail below, Prof. Hauser did engage in research misconduct by intentionally falsifying the research record.

The following facts about the coding are not disputed.

(b) (6), (b) (7)(C) ran the experiment on Cayo Santiago in November and December, 2006, and he performed the first coding. He found that subjects responded slightly more to grammatical stimuli than to ungrammatical (33 out of 63 total responses). The hypothesis of this experiment, as an "expectation violation" study, is that subjects will respond more to ungrammatical stimuli. These results, then, clearly did not support the hypothesis. After receiving them, Prof. Hauser did his own coding of the same trials, as described in the following email:

⁸ To facilitate coding of subject responses for the purposes of this investigation, a copy of the iMovie data was broken up into clips of individual trials by Committee staff.

Subject: rhesus AXA!**Date: March 10, 2007 2:46:36 PM EST**

so, i reanalyzed all the trials from (b) (6), (b) (7)(C) run with AXA on rhesus, using a 2 sec cutoff post-pback, and only counting as YES those responses happening in the 2sec window. it is successful, with wilcoxon at 2.46, $p < .01$ will have (b) (6), (b) (7)(C) recode some of these to make sure!
(Hauser to (b) (6), (b) (7)(C), 3/10/07)

When Prof. Hauser sent (b) (6), (b) (7)(C) trials to recode, he provided the following coding instructions: "remember that we are only counting as yes cases in which animal 1) looks during pback [playback, the running of the auditory stimulus] and maintains look into the 2 sec post-pback period or 2) looks during the 2 sec post-pback period." (Hauser to (b) (6), (b) (7)(C) 3/13/07) At his interview, Prof. Hauser confirmed that this was the correct description of the new protocol: "So to be clear, it should have been either you orient during the sound and keep it oriented after it ends, or you start as soon as it ends, but before 2 seconds." Hauser Interview 5/19/09, at p. 55.

Prof. Hauser describes the "old response criteria" that (b) (6), (b) (7)(C) used to get his non-significant results as follows: "any orientation toward the speaker during or 2 sec after the playback counted as a response." Hauser Response, 5/12/09, p. 37.

Findings and discussion

The Committee finds that there is one distinction between the two criteria: the "old" one would permit a response – a look in the direction of the sound source – that began during the playback period, but did not persist until the end of the playback. In other words, a "quick look" and look away would count under the "old response criteria" but would not count under the criteria of the reanalysis, sent to (b) (6), (b) (7)(C) to use for recoding "to make sure." This conclusion is supported by Prof. Hauser's coding notations on the spreadsheet AXAcodesW/gram/ungram.xls, where he coded as "no" several trials that (b) (6), (b) (7)(C) coded "yes," indicating "turned away before end of pback" or "looked back before end of pback." However, a review of all the instances where Prof. Hauser's codes differ from (b) (6), (b) (7)(C) demonstrates that this difference between the criteria does not account for the change in results from not at all significant (in fact, slightly favoring the result opposing the hypothesis) to significant, with $p < .01$.

Prof. Hauser and (b) (6), (b) (7)(C) coded a total of 201 trials, though some did not count in the final statistical analysis because one subject was excluded for nonresponsiveness. Prof. Hauser's coding differed from (b) (6), (b) (7)(C) for 36 trials; in 29 of the 36 the change from (b) (6), (b) (7)(C) coding was in the direction of significance. Looking at all the trials where Prof. Hauser's coding differed from (b) (6), (b) (7)(C) the Committee finds that there were just four cases where a (b) (6), (b) (7)(C) "yes" was changed to a "no" by Prof. Hauser because of the difference in criteria: Subject 18 Trial 2 (S18 T2); S22 T1; S24 T3 (arguable; look back at instant of playback end); S29 T7.

Of the five instances where Prof. Hauser coded a “no” over (b) (6) (b) (7) (C) “yes” and noted “looked away before end” or a similar comment, in two trials it is clear that the subject did not look away until *after* the end of the playback: S21 T2 and S28 T3. All five trials with these coding notations and a change from a “yes” by (b) (6) (b) (7) (C) to a “no” occur on grammatical stimuli and the changes are towards significance.

On another grammatical trial that Prof. Hauser scored a “no” where (b) (6) (b) (7) (C) scored “yes,” Prof. Hauser noted “after 5 sec” (S29 T5); however the subject’s response begins within two seconds of playback. The Committee finds that these examples of coding changes in the direction of significance where the video contradicts the noted justification support the conclusion that Prof. Hauser was not coding blind to condition, and in fact was focused on condition.

Prof. Hauser’s indifference to his lab’s blind coding protocol is illustrated by the manner in which he had (b) (6) (b) (7) (C) do a recode. When he pasted a list of trials for (b) (6) (b) (7) (C) to code into an email, he did not send him just the subject and trial information; he included the stimuli, so (b) (6) (b) (7) (C) would not be coding blind. Prof. Hauser insists that (b) (6) (b) (7) (C) coded these trials blind: “although he had access to this information, he did not code with this information.” Response 5/12/09, p. 34.

(b) (6) (b) (7) (C) has not verified Prof. Hauser’s claim. At his interview, in fact, he took a different tack; he denied being able to tell from the stimulus what response would support a significant result: “I wouldn’t be able to tell. I mean it’s possible that someone else could, but I wouldn’t know.” (b) (6) (b) (7) (C) Interview 2/10/09, p. 25. (b) (6) (b) (7) (C) did know. He had coded these trials the month before he was asked to recode and reported: “AXA ended up slightly favoring the GRAM, so I’m getting (b) (6) (b) (7) (C) to look it over.” (b) (6) (b) (7) (C) to Hauser, 2/22/07). By the time he ran AXA on Cayo, he had also run other grammar playback experiments with the same stimulus response logic. The Committee finds that his claimed ignorance of the significance of the stimulus column in the following email is not credible:

Subject: coding

Date: March 13, 2007 9:48:00 AM EDT

To: (b) (6) (b) (7) (C)

a few things on AXA. first, here are some trials for you to recode. remember that we are only counting as yes cases in which animal 1) looks during pback and maintains look into the 2 sec post-pback period or 2) looks during the 2 sec post-pback period.

second, once you send these to me, what i need are the number of hab trials per subject, and for each test, what stimuli were played in the test trials.

Thanks

Please Recode these trials:

Subject	trial	stim
1	1	u-A
1	2	g-A
1	3	u-B
1	4	g-B
1	5	u-C
1	6	g-C

3	1	g-D
3	2	u-D
3	3	g-A
4	1	g-D
4	2	u-D
4	3	g-A
4	4	u-A
6	1	u-B
6	2	g-B
6	3	u-C
29	1	u
29	2	g
29	3	u
31	1	u-c
31	2	g-C
31	3	u-D
31	4	g-D
33	1	g-c
33	2	u-c
33	3	g-D

The purpose for (b) (6) (b) (7) (C) coding, as Prof. Hauser indicated, was to "make sure" the reanalysis results were valid (Hauser to (b) (6) (b) (7) (C) 3/10/07). In lieu of having two blind coders and a third coder to serve as tie breaker, Prof. Hauser sometimes chose to have a first coder, and then have a second coder with high inter-coder reliability review a substantial sample of trials as a check. In this case, neither of those coding protocols was operative. When Prof. Hauser coded, he was not a first coder. His reanalysis had turned a patent failure to a success at $p < 0.01$, a fact that he knew, and that should have caused him to proceed with caution. The Committee finds that having his RA review 26 of 202 trials, including only 7 of the 36 trials where his result differed from (b) (6) (b) (7) (C) original result, was not a reasonable means of ensuring the soundness of his work and reconciling the disparity between his and (b) (6) (b) (7) (C) codings. As Prof. Hauser notes, coding 26 trials takes only one to two hours. Hauser Response to IC 5/12/09, p. 38. If Prof. Hauser had genuinely been interested in validating his coding results, he would have invested in the additional few hours to have a complete recoding carried out. The Committee finds that Prof. Hauser's haste to publish successful results on AXA, as shown by the following email messages, led him to intentionally ignore generally accepted practices for analyzing playback response data and protecting against observer bias.

"...I really want to finish up AXA. can you get me the codes soon?" (Hauser to (b) (6) (b) (7) (C), 3/23/07)

...also, what happened with recode of AXA? Are we well correlated? are my analyses good to go? can i write up? would like to move on these soon, so please get in touch about this when you land. (Hauser to (b) (6) (b) (7) (C) 3/29/07)

The Committee finds it particularly troubling that when members of the lab offered to do a recoding, Prof. Hauser resisted. Prof. Hauser has offered the history of the AXA project as an example of the forthright manner in which he deals with disagreements and conflicts in his lab. Hauser CPC Response, 2/8/08 at p. 73. However, Prof. Hauser's

Response to the Investigating Committee revives a misrepresentation he made that was a focal point of the dispute about AXA: he repeatedly refers to (b) (6), (b) (7)(C) first coding as "online" coding. Response, 5/12/09, at p.37. Online codes are "real time," "on the fly" codes recorded by an experimenter as a trial is being run in order to get a sense of how subjects are responding, and they are not considered to be reliable.

(b) (6), (b) (7)(C) discovered the disparity between (b) (6), (b) (7)(C) original coding and Prof. Hauser's (b) (6), (b) (7)(C) suggested that a recoding was appropriate under the circumstances:

While I was looking at the data I noticed there are a good number of differences between the codes in columns B and D (they are 83.7% identical). So we should recode, right? It'll be a great chance for someone else to get familiar with coding these videos since (b) (6), (b) (7)(C) and I are leaving next month and also for us to make a more airtight case... (b) (6), (b) (7)(C) to Hauser (b) (6), (b) (7)(C), 5/23/07)

Prof. Hauser rejected the suggestion that AXA be recoded, based on the false assertion that column B was online:

no. the B column is online. the D column is offline with all yellow rows recoded independently by (b) (6), (b) (7)(C) and i were only off by one trial. so the D column represents the blind coding check with the offline blind coding that i did. (Hauser to (b) (6), (b) (7)(C), 5/23/07)

(b) (6), (b) (7)(C) informed Prof. Hauser that he was going ahead with a full recoding:

...On AXA, given the inconsistencies I got the data from (b) (6), (b) (7)(C), I have already discovered several major errors, so I am just going to go ahead and do a full recode to be sure about things. (b) (6), (b) (7)(C) to Hauser, 5/29/07)

Prof. Hauser expressed his displeasure and disagreed, again repeating that column B was online:

on AXA, i am getting a bit pissed here. there were no inconsistencies! let me repeat what happened. i coded everything. Then (b) (6), (b) (7)(C) coded all the trials highlighted in yellow. we only had one trial that didn't agree. i then mistakenly told (b) (6), (b) (7)(C) to look at column B when he should have looked at column D. B is the online coding. D is the offline coding. so there were no inconsistencies, and this was done as we have always done. but the second point is this: how can you code when you have never watched a rhesus monkey, don't know their behavior, and don't know how we scored? we need to resolve this because i am not sure why we are going in circles. (Hauser to (b) (6), (b) (7)(C), 5/29/07)

(b) (6), (b) (7)(C) rebutted the claim that column B was online code, and pointed out that (b) (6), (b) (7)(C) performance in matching Prof. Hauser's codes 25 out of 26 times should be a cause for concern, not confidence:

I am sorry that this is making you angry, but I respectfully disagree on the AXA data. Column B was not (b) (6), (b) (7)(C) online codes -- they were his offline codes -- this was obvious from the comments embedded in them and he confirmed it when I showed it to him.

I understand that you coded the column D and then he recoded a group that you selected, however given that column B was actually an offline code, and there are large differences between this and D, it seems appropriate to have a proper tie break like we used to do with the tamarin data. Also it doesn't inspire confidence to me that (b) (6), (b) (7)(C) agreed with all but one of the codes you sent him to recode, after all if he was being very careful and blind --- how often does that happen --- that is way beyond any coder reliability rate I have ever seen.

Now, I have been heavily involved in this experiment, I designed the stimuli, and this work is an important part of the story I want to tell with dependency learning, and as such it is very important to me that I can believe it, and right now I just want to double check, because two experienced rhesus coders are showing a very big discrepancy in their data. As a collaborator and coauthor I think this is fair.

While I agree that I do not have much experience with rhesus coding, I have vast experience with tamarin coding and have gone through the criteria with (b) (6), (b) (7)(C) and read your emails to him on coding guidelines. I have tried a few codes and I think I can do a decent job, and I am always happy to discuss any differences. Does this make sense?

(b) (6), (b) (7)(C) to Hauser 5/29/07)

Prof. Hauser "blew up" in reaction to this explanation of why (b) (6), (b) (7)(C) was not willing to rely on Prof. Hauser's coding and (b) (6), (b) (7)(C) check of the coding. That evening, (b) (6), (b) (7)(C) submitted his resignation from the lab:

I have been mulling it over for a while now, and I have decided that it is time for me to move on from the lab.

It has been increasingly clear for a long time now that my interests have been diverging sharply from what the lab does, and it seems like an increasingly inappropriate and uncomfortable place for me.

In any case, this is the best decision for me now, and I hope you can support me in it.

Thanks,

(b) (6), (b) (7)(C) to Hauser, 5/29/07)

(b) (6), (b) (7)(C) followed up the next morning by providing Prof. Hauser a summary of problems with AXA:

Even without being fully expert on the rhesus coding I went ahead and did a recode of the AXA stuff.

Independent of the issue of a third party tie break there are numerous, very big errors in the data.

- 1.) At least 3 (maybe 4, I don't have the data in front of me) of the monkeys had the labels of the U/G switched -- that is listening to the trials it was clear that what had been labelled as a U was a G and vice versa.
- 2.) One whole monkey which had been coded as a test alternating U and G was in fact just hab trials up to the last trial, and has to be excluded.
- 3.) There were several trials which appeared in the data sheet which I could not find anywhere in the videos.
- 4.) Two of the monkeys has the first trial clipped off the video.
- 5.) There were numerous trials which I think should have been bads that were coded, e.g. there was a trial coded where one monkey actually started to fuck another.

In any case, for the majority of trials my codes agreed with both yours and (b) (6), (b) (7)(C) so I guess that the coding is fairly similar. However, it came out miles from significant (.66). With all the mistakes in the original it is hard to even interpret what was going on there.

I am afraid I think that this result is dead in the water, or at the very least the videos need to be completely recoded from scratch corrected, etc.
(b) (6), (b) (7)(C) to Hauser, 5/30/07)

Prof. Hauser did not refute (b) (6), (b) (7)(C) recitation of the problems he found and the conclusion he drew, that the "result is dead in the water." In fact, he concurred, calling AXA a "bust" and initiating plans to rerun it. However, he blamed the "bust" on (b) (6), (b) (7)(C) in an email to (b) (6), (b) (7)(C) that afternoon:

(b) (6), (b) (7)(C) is going to cayo next wed. Given problems of the past, i need to make sure everything is really working on his computer. AXA was a bust. He screwed up the codes. He has lost the last some-B. so, i want him to rerun AXA and want to make sure he is well set up with B before C. so i want to have you run through this with him, do several practice runs and make sure he knows how the output files work.
More in a bit
(Hauser to (b) (6), (b) (7)(C) 5/30/07)

The most striking omission from Prof. Hauser's accounts of AXA project conflict resolution is any mention that the denouement was the departure of (b) (6), (b) (7)(C) from his lab. For the purposes of the Committee's analysis, a telling element in the exchanges set forth above is Prof. Hauser's opposing recoding by telling (b) (6), (b) (7)(C) and (b) (6), (b) (7)(C) that (b) (6), (b) (7)(C) original coding was mere online coding. This dispute occurred a year after the

Syl & Seg and P vs N coding trouble. With those problems as recent history, Prof. Hauser's refusal to check his facts before opposing recoding appears willful. Just three months earlier, he knew (b) (6) (b) (7)(C) coding was offline. This e-mail shows he was waiting for coding to be completed; he was not asking for online coding which would be of minimal interest, and would not be referred to as "analyses":

so, where are we on analyses? did you finish off [f] AXA and pass on to (b) (6) (b) (7)(C)? can you pass files on to me and also tell me the coding criteria? (Hauser to (b) (6) (b) (7)(C) 2/28/07)

The Committee finds that Prof. Hauser did not want a recoding because he wanted to preserve the significant result he had obtained through his reanalysis, and mischaracterizing (b) (6) (b) (7)(C) original results as online coding was a means of resisting the requests of (b) (6) (b) (7)(C). The Committee further finds, based on all the evidence discussed above, that Prof. Hauser did not code the trials blind, and he recorded results that were inconsistent with the video record in order to be able to report a significant result. This constitutes intentional falsification of the research record. One additional basis for this finding is that Prof. Hauser has provided numerous inconsistent accounts of the AXA project: Hauser Response to CPC 9/17/07 at p. 16; Hauser Response to CPC 2/8/08 at p.73; Hauser Response to IC 5/12/09; Hauser Follow-up Response, 5/27/09, p.5. On the question of how the coding was carried out, for example, Prof. Hauser has given the following descriptions:

The RA compiled the data for this experiment and ran a complete coding of the data. I then coded a subset of the trials, and we obtained a high inter-observer reliability. Based on this, we generated a data set and the results suggested a success. I passed this on to one of my graduate students who had helped out with the design and who would be on the paper. Hauser Response to CPC 9/17/07 at p. 16

(b) (6) (b) (7)(C) returned with the data, and having carried out both the online and primary offline coding, I carried out the secondary offline coding. Our initial analyses suggested a success...Based on these analyses, (b) (6) (b) (7)(C) decided to look in greater detail at the pattern of responses. Hauser Response to CPC 2/8/08 at p.73

(b) (6) (b) (7)(C) initially came back from the field and coded the data. As the Committee correctly states, his initial coding revealed no significant difference between responses to grammatical versus ungrammatical test items. ... (b) (6) (b) (7)(C) and I explicitly discussed the details of coding after his initial analyses, and because of this discussion, agreed that a complete recoding was required given the design of the experiment. This is clear in my email (reproduced by the Committee) to (b) (6) (b) (7)(C) where I specify the coding criteria. ...Specifically, in (b) (6) (b) (7)(C) original coding, he included as a "response" any head turn by the subject to the speaker that occurred either during the playback or in the 2 second post-playback period. What we both recognized was that because the detection of a string of sounds as

grammatical (AXA) or not (e.g., AX, AAX) depends upon hearing the entire string, the only meaningful responses arise *after* the entire sound sequence has been played. I am relatively confident that I discussed this coding issue with (b) (6) (b) (7)(C) (b) (6) (b) (7)(C), though it may have been in person as opposed to email. Given my discussion with (b) (6) (b) (7)(C) about the coding procedure, I recoded the entire set of trials, blind to the conditions, and then asked (b) (6) (b) (7)(C) to recode a subset of these trials. Hauser Response to IC 5/12/09, p. 35

When (b) (6) (b) (7)(C) presented the first set of analyses, we both agreed that the coding criteria were wrong. Our motivation for recoding was not to fish for results, but to reduce the noise that would enter by including all responses both during and after playbacks. We then agreed on a coding protocol, and followed the lab's procedures for coding blind. Hauser Response to IC 5/12/09, p. 36

(b) (6) (b) (7)(C) used the old response criteria to do his online scoring: any orientation toward the speaker during or 2 sec after the playback counted as a response. But with the new criteria that (b) (6) (b) (7)(C) and I established, responses were only scored if they occurred within two seconds of the termination of the playback. Hauser Response to IC 5/12/09, p. 37

(b) (6) (b) (7)(C) Just to stick on that point about the criteria though, if you look at page 37 of your response, on number 3, it says, "but with the new criteria, responses were only scored if they occurred within 2 seconds of the termination of the playback."

(b) (6) (b) (7)(C) MH: That's a mistake. But you will see ... actually I repeated the email you gave me which laid out those two response criteria. That's a mistake in there. So to be clear, it should have been either you orient during the sound and keep it oriented after it ends, or you start as soon as it ends, but before 2 seconds.

(b) (6) (b) (7)(C) And either one of those is good?

(b) (6) (b) (7)(C) MH: Exactly.

(b) (6) (b) (7)(C) Hauser Interview, 5/19/09, p.55

(b) (6) (b) (7)(C) No purpose is served in teasing apart the specifics of how the versions differ and speculating as to the possible significance of the variations. Lacking a consistent and credible explanation by Prof. Hauser, the Committee makes its finding of intentional falsification on the basis of the written record, corroborated testimony of witnesses, and reasonable inferences from the undisputed facts found in the record.

§93.313 (f) (3) This research received PHS support. Grant number CM-5-P40RR003640-13 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), provided funds to maintain the rhesus colony on Cayo Santiago. (b) (6) (b) (7)(C) work was supported by NIH National Research Service Award (NRSA) grant F31MH075298.

§93.313 (f) (4) The Committee is not aware of any publications that need correction or retraction as a consequence of this misconduct.

CONFIDENTIAL

Report of Investigating Committee
8 January 2010

§93.313 (f) (5) The Committee finds that Prof. Hauser is the person responsible for the misconduct.

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

The Committee is unaware of any pending applications by Prof. Hauser for federal support.

(b) (6), (b) (7)(C)

(b) (6), (b) (7)(C)

Consideration of Defenses Raised

While the allegations and evidence are detailed in the attached report, we have also carefully considered a number of alternative explanations that would potentially be exculpatory for Prof Hauser. These are described in the following paragraphs.

One general response of Prof. Hauser to our investigation has been to question the motives of the prime complainants (b) (6), (b) (7)(C). Prof. Hauser has suggested that (b) (6), (b) (7)(C) may have acted in concert against him. His view is that for various reasons related to academic rivalry (b) (6), (b) (7)(C) and disgruntlement with their own career (b) (6), (b) (7)(C), these individuals may have been motivated to perpetrate an attack him.

Such collusion, if it had occurred, would therefore be potentially exculpatory and so we took this claim quite seriously. We found that the main (b) (6), (b) (7)(C) complainants did know each other socially, as well as through the Hauser Lab, and they did compare their experiences as they were considering whether to come forward. We did not find however evidence to suggest that they had conspired with the aim of unjustly accusing Prof. Hauser. The Committee has found no basis upon which to conclude that they either fabricated evidence harmful to Prof. Hauser or withheld or destroyed actual evidence that would have helped him. It is clear that some of their concerns about Prof. Hauser appear to have been long standing and predated their association with each other. It is worth noting that (b) (6), (b) (7)(C) was sufficiently troubled by the disappearance of the Theory of Mind videotapes and records in 1999 that she kept Prof. Hauser's e-mail, but there is no evidence that she ever shared that concern with anyone until 2007.

It is also important to realize that these concerns were independently corroborated by an unlikely source: (b) (6), (b) (7)(C). (b) (6), (b) (7)(C) had noted an unexplained change in the data reported in successive drafts of Prof. Hauser's paper. The discrepancy worried (b) (6), (b) (7)(C), and (b) (6), (b) (7)(C) discussed the issue with Dr. Hauser. (b) (6), (b) (7)(C) did not share (b) (6), (b) (7)(C) concern with (b) (6), (b) (7)(C). Nevertheless, (b) (6), (b) (7)(C) account of the events in question was consistent with what (b) (6), (b) (7)(C).

reported, and with the contents of Prof. Hauser's file. It also should be emphasized that (b) (6), (b) (7)(C) was a reluctant witness who spoke only after (b) (6), (b) (7)(C) had been contacted by the Committee.

Likewise, (b) (6), (b) (7)(C) considered, but decided against, reporting (b) (6), (b) (7)(C) experience while (b) (6), (b) (7)(C). (b) (6), (b) (7)(C) the Committee found no evidence of (b) (6), (b) (7)(C) involvement thereafter with any other member of the Hauser Lab.

(b) (6), (b) (7)(C)

A second matter that we considered carefully is that many of the conclusions drawn from the published material related to this case may not actually be incorrect, even if the methods were problematic. Over the course of this investigation Prof. Hauser has frequently cited the replication of the results of many of the projects under investigation as evidence in his favor. We have found that most of the "replications" were not attempts to carry out an exact duplication of the experiments. Rather, they were variations or in some cases entirely different protocols (often with different subject populations). Nevertheless, we do not challenge Prof. Hauser's claim that these subsequent reports provide support for the conclusions he published in the work under investigation. Often these involved claims that non-human primates exhibited or lacked a certain cognitive capacity found in humans, and the argument was that the subsequent studies also showed the same capacity or lack of capacity.

Even so, however, these "replications" do not rebut the allegations of research misconduct, which are not about whether Prof. Hauser's claims about primates are true, but rather about whether he engaged in scientific misconduct in conducting the experiments and in publishing their results. We found credible evidence that, in a series of studies, Prof. Hauser shaded results in the direction of the hypothesis and engaged in other conduct that fails to fulfill a scientist's responsibilities.

That said, we do take the "replications" into account in our concluding remarks on the significance of Prof. Hauser's apparent misconduct, below.

We also carefully considered a third defense. In some of these instances, one result of the re-analysis by students and trainees that revealed the troubling discrepancies was that the results were never published. One might thus conclude that the system had worked just the way it should. It is not unusual for collaborating scientists to spot errors

in each other's work; if these cannot be corrected, abandoning plans to publish the results is often the appropriate response.

We would agree that the challenges, debates, and re-interpretations that routinely take place among collaborating scientists should not occasion an investigation of research misconduct. However, our interpretation of the federal regulations and definitions does not support this defense. For research for which compliance with federal regulations is required, the mere absence of a resulting publication is not in itself exculpatory.

Conclusion

Perhaps we have fulfilled our responsibilities to ORI in the assessments we have offered above of the specific allegations of research misconduct by Prof. Hauser. Nevertheless, we conclude with a brief indication of our own view, as peer faculty, of the significance of our findings within the context of Prof. Hauser's career as a scientist and in comparison with well-known cases of research misconduct by others.

We did not find evidence that Prof. Hauser has been inventing findings out of whole cloth. He did not merely pretend to do research, as some of the most egregious offenders have done; in each case, the problem arose in how Prof. Hauser drew conclusions from his experiments. Moreover, it is entirely possible that in each case Prof. Hauser believed that the results that he wished to report were true. In this sense, one might credit him with a passionate interest in advancing science, in particular with achieving a new appreciation of the dimensions of animal cognition. He has been at the forefront in tracing some of the implications of these putative findings for our understanding of language, morality, and other key functions heretofore regarded as exclusively human. None of the instances of research misconduct that we have examined here necessarily undermines the basis for the high regard in which he is held by many of his colleagues for challenging conventional wisdom and pointing the way to new ways to understand human and animal cognition. This would be particularly the case if, as Prof. Hauser insists, the claims made in the disputed research – whatever his failings in regard to research integrity – are in fact true.

Prof. Hauser's shortcomings in respect to research integrity have in the main consisted instead of repeated instances of cutting corners, of pushing analyses of data further in the direction of significance than the actual findings warranted, and of reporting results as he may have wished them to have been rather than as they actually were. In this regard, it would be no defense to assert that Prof. Hauser may have believed that his claims were true, even though they were not supported by the data from his laboratory. Skepticism above all toward the veracity of one's own hypotheses is of course an essential virtue for scientists, and one that must be modeled for the benefit of trainees.

These remarks on the context of our findings provide no reason to retreat from the findings of fact that we offer here. In some cases, it has been impossible to determine with certainty precisely what happened in experiments and in their interpretation because

CONFIDENTIAL

Report of Investigating Committee
8 January 2010

of the lack of records. While we cannot rule out innocent explanations in these cases, our review has found that the preponderance of the evidence has not supported them. In our investigation, we found evidence that Prof. Hauser repeatedly valued the primacy and impact of his ideas above an accurate representation of his scientific methods and the integrity of the data obtained to support them.

Respectfully submitted,

(b) (6), (b) (7)(C)



