MATHEMATICA
Policy Research, Inc.

MEMORANDUM

600 Maryland Ave., SW, Suite 550
Washington, DC 20024-2512
Telephone (202) 484-9220
Fax (202) 863-1763
www.mathematica-mpr.com

TO: Sophie van der Bijl

FROM: Matt Sloan **DATE:** 8/5/2009

SUBJECT: Burkina Faso Revised Data Delivery

Thank you very much to you and your colleagues at MCC for providing us with comments on our draft data set and documentation for the BRIGHT Impact Evaluation. Please find attached the final version of the data set and documentation incorporating revisions we made in response to the suggested changes. To facilitate the process of reviewing how we responded to the changes, we have included below a response to each of the comments you made. Please do not hesitate to contact us if you have any questions or need further information.

- Interviewer names could be coded to an ID number.
  - Since this variable (hc3) is not useful for analysis, it was dropped from the data file.
- Names of heads of households should be eliminated from the data file.
  - Hc8 (name of head of household) was dropped from the data file.
- Latitude and longitude coordinates should be aggregated up to the village level (at least), if this has not been done already. Or we may need to go even further on the GIS data (latitude and longitude), since that essentially identifies each village, which is otherwise not identified.
  - All GIS variables were dropped (hc13lond, hc13lonm, hc13latd, hc13latm, eclond, eclonm, eclatd, and eclatm). Village is identified by a unique ID number (hc1).
  - In addition, after reviewing the data file further, we dropped village name (village), student's village (vilelev) and school name (ecoleid).

- In the Codebook, why the odd frequency value for variable hl5 (child age) (p16), where the proportion of children aged 11 are substantially fewer? Is this a flaw in the data, or a quirk of reporting in the summary (e.g., if rounding to 10 or even numbers)?

  - After investigating this, we were unable to identify a problem with data file or the procedures used to collect the data. We conclude that the lower than average frequency of children reported to be 11 years old could reasonably be attributed to natural variation. If we expected children's age to be evenly distributed, we would expect to see 12.5% in each group. While those reported to be aged 11 appear somewhat lower than this expectation, those aged 5 and 7 are somewhat higher. This does not rule out rounding errors, misreporting or other problems, but we have no evidence that would suggest that this systematically led to underreporting for those aged 11 years (or over-reporting for other ages). As a result, no change to the data file or documentation was made.

Variable Name:	hl5
Label:	Child Age
Question Text:	How old is (name)? How old was (name) on his/her last birthday?
Universe:	All
N:	21,747

Frequency	Value	Percent	Cum. Percent
3,713	5	17.07	17.07
2,428	6	11.16	28.24
3,652	7	16.79	45.03
2,774	8	12.76	57.79
2,704	9	12.43	70.22
2,546	10	11.71	81.93
1,242	11	5.71	87.64
2,688	12	12.36	100.00

- On school distance (ed7) and school travel time (ed8) variables, what are the units?

  - The units for distance are kilometers and the units for time are minutes. We have revised the variable labels and documentation to reflect this.

- For variables sc4_lgi . . . sc4_6fr, how were the two questions included in a single response?

  - For this series of questions, school administrators were asked how many boys and girls were enrolled in each grade and then how many boys and girls were repeating the grade. The documentation failed to make these distinctions. We have revised the documentation and variables labels to accurately reflect them.

- Why does the range of response for schools on variable sc9 not match the range of response for the same question to students (ed16)?

  - These questions ask both the household and school to report on feeding programs. The interviewers were provided with likely responses (e.g. breakfast, snack, lunch) and an "other" category. While in the field, interviewers added additional information in the "other" in the field in an inconsistent manner which was preserved in the data file. Other differences are likely due to the head of household being unaware of the specifics of the feeding program. We have added a footnote with this explanation on page 43 of the codebook.

- The descriptive labels for hadschool_1, hadschool_2, and hadschool_3 should be corrected to state that there was a record of children attending school in the village—not that there were no children attending school, as the description now states.

  - The variable labels have been revised in the documentation and data file.

- The Manual never addressed or dismissed the possibility that the survey introduced a bias. The Users' Manual section on "School Sampling" (p2) states that, "Interviewers then selected the up to three schools closest to the village center, within 10 kilometers, as the schools to be surveyed for that village." For example, perhaps "elite" schools are located in the village center—or conversely located beyond the village where there are fewer distractions. What if religious schools are centrally located near mosques or churches, with public schools located further from villages?

  - The bias mentioned is possible but we believe is extremely unlikely. This is due to the fact that the great majority of villages had no school or one school, and most kids traveled to the school closest to home. We have added a footnote on page 2 of the user's manual to explain this.