Online Appendix to:

How to get better survey data more efficiently

Mollie J. Cohen[*] and Zach Warner[†]

February 2, 2019

[*] Assistant Professor, University of Georgia. email: mj.cohen@uga.edu, web: http://www.molliecohen.com.

[†] Postdoctoral Research Fellow, Cardiff University. email: WarnerZ@cardiff.ac.uk, web: http://www.zachwarner.net.

Quality control procedures

The following list gives each quality control procedure name, description, and coding rules.

Auditor problems, missingness:

1. no_respondent_auditor: describes whether an auditor noticed that there was no respondent (from silent audio capture). Coded as 1 if no respondent heard and 0 otherwise.

2. no_gps_auditor: describes whether an auditor noticed that there was no GPS data available for the interview (from silent GPS capture). Coded as 1 if no GPS data available and 0 otherwise.

3. no_camera_auditor: describes whether an auditor noticed that there was no image captures uploaded with an interview (from various image captures). Coded as 1 if no images available and 0 otherwise.

Auditor problems, consent form:

4. consent_wrongtime_auditor: describes whether an auditor noticed the enumerator mis-stating the expected duration of the survey when asking the respondent for informed consent (from silent audio capture). Coded as 1 if the stated time is incorrect and 0 otherwise.

5. consent_notgiven_auditor: describes whether an auditor began the survey despite the respondent not giving informed consent (from silent audio capture). Coded as 1 if informed consent was not given and 0 otherwise.

6. consent_notread_auditor: describes whether an auditor began the survey without reading the consent form (from silent audio capture). Coded as 1 if the consent form was not read and 0 otherwise.

7. consent_partread_auditor: describes whether an auditor began the survey after only partially reading the consent form (from silent audio capture). Coded as 1 if the consent form was partially read and 0 otherwise.

8. consent_misread_auditor: describes whether an auditor misread the consent form (from silent audio capture). Coded as 1 if the consent form was misread and 0 otherwise.

Auditor problems, improper people involved:

9. friend_interviewed_auditor: describes whether an auditor noticed that the respondent appeared to know the enumerator personally (from silent audio capture). Coded as 1 if the enumerator knew the respondent and 0 otherwise.

10. wrong_picture_auditor: describes whether an auditor noticed that the front-facing image uploaded with an interview captured a face that was not the enumerator's (from image captures). Coded as 1 if the face photographed was not that of the enumerator and 0 otherwise.

11. wrong_voice_auditor: describes whether an auditor noticed that the voice conducting the interview was not that of the enumerator (from silent audio capture). Coded as 1 if the enumerator's voice was incorrect and 0 otherwise.

12. self_interview_auditor: describes whether an auditor noticed that the enumerator interviewed him- or herself instead of a respondent (from silent audio capture). Coded as 1 if the enumerator interviewed him- or herself and 0 otherwise.

13. interviewer_interview_auditor: describes whether an auditor noticed that an enumerator interviewed another enumerator instead of a respondent (from silent audio capture). Coded as 1 if the enumerator interviewed another enumerator and 0 otherwise.

Auditor problems, suspicious behavior:

14. interviewer_abandon_auditor: describes whether an auditor noticed that the enumerator abandoned the interview for any reason (from audio and image captures, as well as the interview log). Coded as 1 if the enumerator abandoned the interview and 0 otherwise.

15. interviewee_abandon_auditor: describes whether an auditor noticed that the respondent abandoned the interview for any reason (from audio and image captures, as well as the interview log). Coded as 1 if the interviewee abandoned the interview and 0 otherwise.

16. wrong_location_auditor: describes whether an auditor noticed that the interview took place in a proscribed location, such as a supermarket, gas station, or university (from audio, image, and GPS captures, as well as contextual clues). Coded as 1 if the interview was conducted in a proscribed location and 0 otherwise.

17. location_moved_auditor: describes whether an auditor noticed that the interview began and ended in different locations (from audio, image, and GPS captures as well as contextual clues). Coded as 1 if the starting and ending locations were not the same and 0 otherwise.

18. attempts_exhausted_auditor: describes whether the auditor noticed that the interview uploaded had exhausted all attempts to interview without successfully completing an interview (from the logs). Coded as 1 if all attempts were exhausted and no survey questions were answered and 0 otherwise.

19. tooshort_toolong_auditor: describes whether the auditor noticed (from the log) that the interview was completed too quickly, or took too long to complete, based on country-specific thresholds (but typically less than 25 minutes or more than 2 hours, respectively). Coded as 1 if the interview was either too short or too long and 0 otherwise.

20. airplane_mode_auditor: describes whether the auditor noticed that the CAPI device was switched to airplane mode to prevent a network connection (from the log). Coded as 1 if the device was in airplane mode and 0 otherwise.

Auditor problems, misreading:

21. interviewer_gaveopinion_auditor: describes whether an auditor noticed that the enumerator gave his or her opinion on a survey question or answer to the respondent (from a silent audio capture). Coded as 1 if the enumerator gave his or her opinion and 0 otherwise.

22. question_oneinterpreted_auditor: describes whether an auditor noticed that the enumerator interpreted one survey question for the respondent (from a silent audio capture). Coded as 1 if exactly one question was interpreted and 0 otherwise.

23. question_twointerpreted_auditor: describes whether an auditor noticed that the enumerator interpreted two survey questions for the respondent, from a silent audio capture (from a silent audio capture). Coded as 1 if exactly two questions were interpreted and 0 otherwise.

24. question_manyinterpreted_auditor: This variable describes whether an auditor noticed that the enumerator interpreted three or more survey questions for the respondent (from a silent audio capture). Coded as 1 if three or more questions were interpreted and 0 otherwise.

25. question_oneskip_auditor: describes whether an auditor noticed that the enumerator skipped a survey question (from a silent audio capture). Coded as 1 if exactly one question was skipped and 0 otherwise.

26. question_twoskip_auditor: describes whether an auditor noticed that the enumerator skipped two survey questions (from a silent audio capture). Coded as 1 if exactly two questions were skipped and 0 otherwise.

27. question_manyskip_auditor: describes whether an auditor noticed that the enumerator skipped three or more survey questions (from a silent audio capture). Coded as 1 if three or more questions were skipped and 0 otherwise.

28. question_onemisread_auditor: describes whether an auditor noticed that the enumerator misread a survey question (from a silent audio capture). Coded as 1 if exactly one question was misread and 0 otherwise.

29. question_twomisread_auditor: describes whether an auditor noticed that the enumerator misread two survey questions (from a silent audio capture). Coded as 1 if exactly two questions were misread and 0 otherwise.

30. question_manymisread_auditor: describes whether an auditor noticed that the enumerator misread three or more survey questions (from a silent audio capture). Coded as 1 if three or more questions were misread and 0 otherwise.

31. question_onefast_auditor: describes whether an auditor noticed that the enumerator read a survey question too fast (from a silent audio capture). Coded as 1 if exactly one question was read too fast and 0 otherwise.

32. question_twofast_auditor: describes whether an auditor noticed that the enumerator read two survey questions too fast (from a silent audio capture). Coded as 1 if exactly two questions were read too fast and 0 otherwise.

33. question_manyfast_auditor: describes whether an auditor noticed that the enumerator read three or more survey questions too fast (from a silent audio capture). Coded as 1 if three or more questions were read too fast and 0 otherwise.

Auditor problems, other:

34. other_qac_abandoned_security_auditor: describes whether the auditor abandoned the interview due to security concerns for him- or herself. Coded as 1 if the auditor abandoned the interview due to security concerns and 0 otherwise.

35. other_qac_potential_fraud_auditor: describes whether the auditor noticed anything in reviewing the interview that suggested potential enumerator fraud. Coded as 1 if the auditor suspected fraud and 0 otherwise.

36. other_qac_potential_fraud_loc_auditor: describes whether the auditor noticed anything in reviewing the interview that suggested potential enumerator fraud based on the location of the interview. Coded as 1 if the auditor suspected fraud due to location and 0 otherwise.

37. other_qac_potential_fraud_quota_auditor: describes whether the auditor noticed anything in reviewing the interview that suggested potential enumerator fraud due to demographic quota issues. Coded as 1 if the auditor suspected fraud for quota reasons and 0 otherwise.

38. other_qac_potential_fraud_many_ppl_auditor: describes whether the auditor noticed anything in reviewing the interview that suggested potential enumerator fraud due to multiple apparent respondents in the same interview. Coded as 1 if the auditor suspected fraud for this reason and 0 otherwise.

39. other_qac_potential_fraud_other_auditor: describes whether the auditor noticed anything in reviewing the interview that suggested potential enumerator fraud for reasons not otherwise noted. Coded as 1 if the auditor suspected fraud and 0 otherwise.

40. other_qac_third_party_influence_auditor: describes whether the auditor noticed anything in reviewing the interview that suggested the respondent was unduly influenced by a third party present during the interview. Coded as 1 if the auditor suspected such influence and 0 otherwise.

41. other_qac_incomplete_read_auditor: describes whether the auditor noticed any place where the enumerator did not completely read an item or response (not noted elsewhere). Coded as 1 if the auditor noticed such incompletes and 0 otherwise.

42. other_qac_duration_problem_auditor: describes whether the auditor noticed any other problem with the duration of the interview (not noted elsewhere). Coded as 1 if the auditor noted such a problem and 0 otherwise.

43. other_qac_read_wrong_inc_auditor: describes whether the auditor noticed the enumerator reading wrong response options for a question (not noted elsewhere). Coded as 1 if the enumerator misread response options and 0 otherwise.

44. other_qac_missing_attachments_auditor: describes whether the auditor noticed any other uploads missing (not noted elsewhere). Coded as 1 if the attachments were missing and 0 otherwise.

45. other_qac_record_wrong_ans_auditor: describes whether the auditor mis-recorded an answer given by the respondent (from a silent audio capture). Coded as 1 if the auditor mis-recorded a response and 0 otherwise.

46. other_qac_technical_error_auditor: describes whether the auditor noticed a the enumerator make a technological error during the interview. Coded as 1 if the enumerator commited a technological error and 0 otherwise.

47. other_qac_no_consent_heard_auditor: describes whether the auditor could not hear or understand whether the respondent gave informed consent (from a silent audio capture) Coded as 1 if the consent could not be heard or understood and 0 otherwise.

48. other_qac_ambient_noise_loud_auditor: describes whether the auditor noticed loud ambient noises making it difficult to hear the voices of the enumerator and/or respondent (from a silent audio capture). Coded as 1 if ambient noise was excessive and 0 otherwise.

49. other_qac_outside_geo_area_auditor: describes whether the auditor noticed that the enumerator conducted the interview outside of the assigned geofence for that interview (from GPS captures). Coded as 1 if the interview was outside the geofence and 0 otherwise.

50. other_qac_tech_problems_auditor: describes whether the auditor noticed a technological problem during the interview. Coded as 1 if there was a technological problem and 0 otherwise.

51. other_qac_one_geocerca_auditor: describes whether the auditor noticed that the interview was conducted in the wrong geofence assigned to that interview. Coded as 1 if the interview was conducted in the wrong geofence and 0 otherwise.

52. other_qac_interpreter_used_auditor: describes whether the auditor noticed that the enumerator relied on an intermediary to translate communications to or from the respondent. Coded as 1 if the respondent used an interpreter and 0 otherwise.

53. other_qac_age_quota_unmet_auditor: describes whether the auditor noticed that the respondent did not meet the age quota. Coded as 1 if the respondent did not meet the age quota and 0 otherwise.

54. other_qac_wrong_gender_quota_auditor: describes whether the auditor noticed that the respondent did not meet the gender quota. Coded as 1 if the respondent did not meet the gender quota and 0 otherwise.

55. other_qac_cant_speak_lang_auditor: describes whether the auditor noticed that the respondent could not speak the language of enumeration. Coded as 1 if the respondent did not speak the language of the survey and 0 otherwise.

56. other_qac_resp_read_questionnaire_auditor: describes whether the auditor noticed that the respondent read the questionnaire directly. Coded as 1 if the respondent read the questionnaire and 0 otherwise.

57. other_qac_quota_problem_auditor: describes whether the auditor noticed a problem with the interview with respect to fulfilling demographic quotas. Coded as 1 if there was such a problem and 0 otherwise.

58. other_qac_comment_error_auditor: describes whether the auditor noticed a miscellaneous enumerator error (not otherwise noted) Coded as 1 if there was such an error and 0 otherwise.

59. other_qac_bad_etiquette_auditor: describes whether the auditor noticed the enumerator displaying bad etiquette, such as answering his or her phone. Coded as 1 if there the enumerator displayed bad etiquette and 0 otherwise.

60. other_qac_consent_error_auditor: describes whether the auditor noticed a miscellaneous problem with the consent form (not otherwise noted). Coded as 1 if there was such a consent error and 0 otherwise.

61. other_qac_consent_read_wrong_auditor: describes whether the auditor noticed the enumerator make a miscellaneous error reading the consent form (not otherwise noted). Coded as 1 if such an error was made and 0 otherwise.

62. other_qac_consent_request_id_auditor: describes whether the auditor noticed the enumerator request identifying information from the respondent when asking for informed consent. Coded as 1 if such a request was made and 0 otherwise.

63. other_qac_other_error_auditor: describes whether the auditor noticed any enumerator errors not otherwise noted. Coded as 1 if there was such an error and 0 otherwise.

64. qac_otherproblem_auditor: describes whether the auditor noticed any problems not otherwise noted. Coded as 1 if there was such a problem and 0 otherwise.

STG flags:

65. early_termination_flag: describes whether STG flagged the interview as being terminated before completion, as indicated by the enumerator pressing the "early termination" button. Coded as 1 if the interview was terminated early and 0 otherwise.

66. set_as_complete_flag: describes whether STG flagged the interview as being manually set to "complete," as opposed to automatically set to complete after the interview concluded. Coded as 1 if the interview was manually set to complete and 0 otherwise.

67. gender_consistency_flag: describes whether STG flagged the interview for having gender set to different values by the enumerator at the beginning and end of the interview. Coded as 1 if the gender was inconsistent throughout the interview to complete and 0 otherwise.

68. stopped_continued_flag: describes whether STG flagged the interview as stopping and restarting. Coded as 1 if the interview stopped and restarted and 0 otherwise.

69. uploaded_by_another_flag: describes whether STG flagged the interview as being uploaded to the server by an enumerator other than the one who conducted the interview. Coded as 1 if the interview was conducted and uploaded by different enumerators and 0 otherwise.

70. no_gps_continue_flag: describes whether STG flagged the interview as having no GPS data but the enumerator manually choosing to continue anyway. Coded as 1 if the enumerator manually over-rode the "lack of GPS data" warning and continued with the interview, and 0 otherwise.

71. outside_geofence_continue_flag: describes whether STG flagged the interview as being outside the geofence assigned to that interview but the enumerator manually chose to continue anyway. Coded as 1 if the enumerator manually over-rode the "outside of geofence" warning and continued with the interview, and 0 otherwise.

72. outside_geofence_cancel_flag: describes whether STG flagged the interview as being outside the geofence assigned to that interview and the enumerator manually chose to cancel it. Coded as 1 if the enumerator manually canceled the interview for this reason, and 0 otherwise.

73. silent_attachments_flag: describes whether STG flagged any audio captures as being completely silent. Coded as 1 if the interview attachments include at least one silent audio capture and 0 otherwise.

74. stop_without_save_flag: describes whether STG flagged the interview as being stopped without saving. Coded as 1 if the interview stopped without saving and 0 otherwise.

75. version_changed_flag: describes whether STG flagged the survey version as changing during the interview. Coded as 1 if the survey version changed over the course of the interview and 0 otherwise.

R scripts, completion and Percentmatch:

76. completion_pc_script: describes the completion percentage for the interview, generated automatically via an R script. Coded as a numeric bounded between 0 and 1 and computed as the proportion of substantive questions (i.e., survey items) to which the respondent gave a valid answer.

77. enumerator_comp_pc_script: describes the mean completion percentage across all interviews conducted by the enumerator, generated automatically via an R script. Coded as a numeric bounded between 0 and 1 and computed as the mean proportion of substantive questions (i.e., survey items) to which the respondent gave a valid answer for all interviews conducted by that enumerator.

78. enumerator_comp_pc_ur_gap_script: describes the (absolute-valued) gap between the mean completion percentage across all interviews conducted by the enumerator in urban versus rural primary sampling units, generated automatically via an R script. Coded as a numeric bounded between 0 and 1 and computed as the mean proportion of substantive questions (i.e., survey items) to which the respondent gave a valid answer for all interviews conducted by that enumerator.

79. top_pc_match_script: describes the highest Percentmatch value for the interview (i.e., the maximum overlap with any other interview), generated automatically via an R script. Coded as a numeric bounded between 0 and 1, computed as the maximum proportion of substantive questions (i.e., survey items) to which the respondent's response is identical to those of another interview.

80. pc_match_top_decile_script: describes whether the interview's top Percentmatch value is in the top decile of all Percentmatch values in the data, generated automatically via an R script. Coded as 1 if the interview is in the top decile and 0 otherwise.

81. pc_match_bot_decile_script: describes whether the interview's top Percentmatch value is in the bottom decile of all Percentmatch values in the data, generated automatically via an R script. Coded as 1 if the interview is in the bottom decile and 0 otherwise.

R scripts, participation rates:

82. enumerator_noh_pc_script: describes the proportion of interview attempts made by the enumerator that are marked as "no one home," generated automatically via an R script. Coded as a numeric bounded between 0 and 1.

83. enumerator_int_pc_script: describes the proportion of interview attempts made by the enumerator that resulted in successful interviews, generated automatically via an R script. Coded as a numeric bounded between 0 and 1.

84. enumerator_ref_pc_script: describes the proportion of interview attempts made by the enumerator that are marked as "refusal," generated automatically via an R script. Coded as a numeric bounded between 0 and 1.

85. enumerator_noh_pc_ur_gap_script: describes the (absolute-valued) difference in proportions of interview attempts made by the enumerator that are marked as "no one home" between urban and rural primary sampling units, generated automatically via an R script. Coded as a numeric bounded between 0 and 1.

86. enumerator_int_pc_ur_gap_script: describes the (absolute-valued) difference in proportions of interview attempts made by the enumerator that resulted in successful interviews between urban and rural primary sampling units, generated automatically via an R script. Coded as a numeric bounded between 0 and 1.

87. enumerator_ref_pc_ur_gap_script: describes the (absolute-valued) difference in proportions of interview attempts made by the enumerator that are marked as "refusal" between urban and rural primary sampling units, generated automatically via an R script. Coded as a numeric bounded between 0 and 1.

R scripts, cluster sampling:

88. cluster_too_big_script: describes whether the sampling cluster contained more interviews than fieldwork protocols requires, generated automatically via an R script. Coded as 1 if 10 or more interviews were in the cluster and 0 otherwise.

89. cluster_too_small_script: describes whether the sampling cluster contained fewer interviews than fieldwork protocols requires, generated automatically via an R script. Coded as 1 if 1 or fewer interviews were in the cluster and 0 otherwise.

90. no_cluster_geo_variation_script: describes whether there was any variation in the GPS coordinates across all interviews in the sampling cluster, generated automatically via an R script. Coded as 1 if there was no geographic variation across interviews in the cluster and 0 otherwise.

91. no_upm_geo_variation_script: describes whether there was any variation in the GPS coordinates across all interviews in the primary sampling unit, generated automatically via an R script. Coded as 1 if there was no geographic variation across interviews in the PSU and 0 otherwise.

92. no_upm_cluster_variation_script: describes whether there was any variation in the unique cluster identification numbers across all interviews in the primary sampling unit, generated automatically via an R script. Coded as 1 if there was no cluster variation across interviews in the PSU and 0 otherwise.

93. cluster_dispersion_script: describes the compactness and separation of clusters within primary sampling units, generated automatically via an R script. Coded as numeric, bounded between -1 and 1, computed using the global average silhouette within a primary sampling unit.

94. cluster_disp_top_decile_script: describes whether the interview's cluster dispersion (silhouette) value is in the top decile of all cluster dispersion values in the data, generated automatically via an R script. Coded as 1 if the interview is in the top decile and 0 otherwise.

95. cluster_disp_bot_decile_script: describes whether the interview's cluster dispersion (silhouette) value is in the bottom decile of all cluster dispersion values in the data, generated automatically via an R script. Coded as 1 if the interview is in the bottom decile and 0 otherwise.

96. cluster_disp_other_script: describes whether there were any other problems measuring cluster dispersion (silhouette) value, generated automatically via an

R script. Coded as 1 if the interview encountered any error computing cluster dispersion not otherwise noted, 0 otherwise.

R scripts, timing:

97. duration_script: describes the absolute duration of the interview, generated automatically via an R script. Coded as numeric, in seconds, strictly positive and integer-valued.

98. netduration_script: describes the duration of the interview, net of screening questions, generated automatically via an R script. Coded as numeric, in seconds, strictly positive and integer-valued.

99. duration_diff_script: describes the difference in duration and net duration of the interview, generated automatically via an R script. Coded as numeric, in seconds, integer-valued.

100. short_ave_question_time_script: describes whether the average time between question prompts was too short, generated automatically via an R script. Coded as 1 if the mean question time across the interview is less than 5 seconds and 0 otherwise.

101. short_ave_attempt_time_script: describes whether the average time between interview attempts was too short, generated automatically via an R script. Coded as 1 if the mean time between attempted interviews is less than 5 seconds and 0 otherwise.

102. long_run_time_script: describes whether the total runtime of the interview was too long, generated automatically via an R script. Coded as 1 if the total runtime for the interview exceeds three hours and 0 otherwise.

103. big_time_jump_script: describes whether there are any large time jumps between questions in the interview log, generated automatically via an R script. Coded as 1 if any time jumps between questions exceeded 10 minutes and 0 otherwise.

104. time_goes_back_script: describes whether there are any backward time jumps in the interview log, generated automatically via an R script. Coded as 1 if any backward time jumps occurred and 0 otherwise.

105. time_out_of_bounds_script: describes whether there are any timestamps in the interview log that are outside the dates over which fieldwork was conducted, generated automatically via an R script. Coded as 1 if any timestamps are out of fieldwork dates and 0 otherwise.

R scripts, network connectivity:

106. lss_gps_disabled_script: describes whether GPS location service was set as "disabled" by the enumerator, generated automatically via an R script. Coded as 1 if GPS location service was disabled for the interview and 0 otherwise.

107. lss_net_disabled_script: describes whether network location service was set as "disabled" by the enumerator, generated automatically via an R script. Coded as 1 if network location service was disabled for the interview and 0 otherwise.

108. mobile_disabled_script: describes whether mobile data was disabled by the enumerator, generated automatically via an R script. Coded as 1 if mobile data was off for the interview and 0 otherwise.

109. use_gps_altered_script: describes whether the "use GPS" setting was set as "off" by the enumerator, generated automatically via an R script. Coded as 1 if "use GPS" was "off" for the interview and 0 otherwise.

110. real_gps_altered_script: describes whether the "use real GPS only" setting was set as "on" by the enumerator, generated automatically via an R script. Coded as 1 if "use real GPS only" was "on" for the interview and 0 otherwise.

111. no_lss_gps_captures_script: describes whether GPS location services settings were captured during the interview, generated automatically via an R script. Coded as 1 if no GPS location service settings were captured and 0 otherwise.

112. no_lss_net_captures_script: describes whether network location services settings were captured during the interview, generated automatically via an R script. Coded as 1 if no network location service settings were captured and 0 otherwise.

113. no_mobile_captures_script: describes whether mobile data settings were captured during the interview, generated automatically via an R script. Coded as 1 if no mobile data settings were captured and 0 otherwise.

114. no_use_gps_captures_script: describes whether any "use GPS" settings were captured during the interview, generated automatically via an R script. Coded as 1 if no use GPS settings were captured and 0 otherwise.

115. no_use_real_gps_captures_script: describes whether any real GPS coordinates (as opposed to approximate coordinates triangulated by WiFi or mobile connections) were captured during the interview, generated automatically via an R script. Coded as 1 if no real GPS coordinates were captured and 0 otherwise.

116. multiple_lss_gps_captures_script: describes whether multiple GPS location services settings were logged during the interview, generated automatically via an R script. Coded as 1 if GPS location service settings were logged more than once and 0 otherwise.

117. multiple_lss_net_captures_script: describes whether multiple network location services settings were logged during the interview, generated automatically via an R script. Coded as 1 if network location service settings were logged more than once and 0 otherwise.

118. multiple_mobile_captures_script: describes whether multiple mobile data settings were logged during the interview, generated automatically via an R script. Coded as 1 if mobile data settings were logged more than once and 0 otherwise.

119. multiple_use_gps_captures_script: describes whether multiple "use GPS" settings were logged during the interview, generated automatically via an R script. Coded as 1 if "use GPS" settings were logged more than once and 0 otherwise.

120. multiple_use_real_gps_captures_script: describes whether multiple "use real GPS only" settings were logged during the interview, generated automatically via an R script. Coded as 1 if "real GPS" settings were logged more than once and 0 otherwise.

R scripts, quotas:

121. age_script: describes the age of the respondent, generated automatically via an R script. Coded as numeric, counted in years (integer-valued and strictly positive).

122. age_quota_invalid_script: describes whether the age given was invalid for the quota, generated automatically via an R script. Coded as 1 if the respondent falls into an invalid age quota category, and 0 otherwise.

123. female_script: describes the stated gender of the respondent, generated automatically via an R script. Coded as 1 if the respondent is female and 0 otherwise.

124. age_quota_young_script: describes the age category of the respondent, generated automatically via an R script. Coded as 1 if the respondent falls into the (country-specific) quota category for the youngest third of respondents and 0 otherwise.

125. age_quota_middle_script: describes the age category of the respondent, generated automatically via an R script. Coded as 1 if the respondent falls into the (country-specific) quota category for the middle third of respondents and 0 otherwise.

126. age_quota_old_script: describes the age category of the respondent, generated automatically via an R script. Coded as 1 if the respondent falls into the (country-specific) quota category for the oldest third of respondents and 0 otherwise.

127. female_age_quota_young_script: describes the age-gender category of the respondent, generated automatically via an R script. Coded as 1 if the respondent falls into the (country-specific) quota category for the youngest third of respondents and is female, and 0 otherwise.

128. female_age_quota_middle_script: describes the age-gender category of the respondent, generated automatically via an R script. Coded as 1 if the respondent falls into the (country-specific) quota category for the middle third of respondents and is female, and 0 otherwise.

129. female_age_quota_old_script: describes the age-gender category of the respondent, generated automatically via an R script. Coded as 1 if the respondent falls into the (country-specific) quota category for the oldest third of respondents and is female, and 0 otherwise.

R scripts, comments:

130. enumerator_comment_script: describes whether the enumerator conducting the interview made any comments in the interview log, generated automatically via an R script. Coded as 1 if the enumerator made any comment and 0 otherwise.

131. qc_comment_script: describes whether a fieldwork supervisor made any comments in the interview log, generated automatically via an R script. Coded as 1 if the supervisor made any comment and 0 otherwise.

132. reviewer_comment_script: describes whether an auditor made any comments in the interview log, generated automatically via an R script. Coded as 1 if the auditor made any comment and 0 otherwise.

133. any_comment_script: describes whether any project staff made any comments in the interview log, generated automatically via an R script. Coded as 1 if any staff made any comment and 0 otherwise.

R scripts, location:

134. no_gps_captures_script: describes whether no GPS data were captured during the interview, generated automatically via an R script. Coded as 1 if no GPS data are available and 0 otherwise.

135. little_gps_change_script: describes if many attempted interviews were made with very little variation in GPS captures, generated automatically via an R script. Coded as 1 if the ratio of attempts to unique GPS locations captured during the interview is greater than five and 0 otherwise.

136. attempts_no_gps_change_script: describes if many consecutive attempted interviews were made without a change in location, generated automatically via an R script. Coded as 1 if seven or more consecutive attempted interviews were made at the same GPS coordinates and 0 otherwise.

137. big_gps_jump_script: describes if large jumps in GPS coordinates were captured between attempted interviews, generated automatically via an R script. Coded as 1 if any jumps of 10 kilometers or more are observed between attempts and 0 otherwise.

R scripts, miscellaneous:

138. broken_photo_script: describes whether the front-facing image capture of the enumerator was missing, broken, or unable to be processed for quality, generated automatically via an R script. Coded as 1 if the image was unreadable and 0 otherwise.

139. bad_photo_script: describes whether the front-facing image capture of the enumerator contained very little pixel variation, generated automatically via an R script. Coded as 1 if the variation in pixel color was less than 0.1 on a scale from 0 to 1 and 0 otherwise.

140. device_battery_script: describes the device battery percentage at the time the interview began, generated automatically via an R script. Coded as a numeric value bounded between 0 and 1.

141. ur: describes whether the interview was conducted in an urban or rural primary sampling unit. Coded as 1 if urban and 0 if rural.

142. enumerator_id_script: uniquely valid enumerator identification numbers, generated automatically via an R script. Coded as a factor.

143. device_id_script: uniquely valid CAPI device identification numbers, generated automatically via an R script. Coded as a factor.

Models studied

The following list gives the caret name and description for each model studied in our classification task. We chose these models for their diversity of underlying approach and computational stability.

1. avNNet: Model Averaged Neural Network
2. bagEarthGCV: Bagged MARS using gCV Pruning
3. bagFDAGCV: Bagged Flexible Discriminant Analysis using gCV Pruning
4. bayesglm: Bayesian Generalized Linear Model
5. C5.0: C5.0
6. C5.0Rules: Single C5.0 Ruleset
7. C5.0Tree: Single C5.0 Tree
8. earth: Multivariate Adaptive Regression Spline
9. fda: Flexible Discriminant Analysis
10. gamSpline: Generalized Additive Model using Splines
11. gcvEarth: Multivariate Adaptive Regression Splines
12. glm: Generalized Linear Model
13. glmnet: glmnet
14. hdda: High Dimensional Discriminant Analysis
15. hdrda: High-Dimensional Regularized Discriminant Analysis
16. multinom: Penalized Multinomial Regression
17. naive_bayes: Naive Bayes
18. nb: Naive Bayes
19. nnet: Neural Network
20. pam: Nearest Shrunken Centroids

21. parRF: Parallel Random Forest
22. pcaNNet: Neural Networks with Feature Extraction
23. pda: Penalized Discriminant Analysis
24. pda2: Penalized Discriminant Analysis
25. rbfDDA: Radial Basis Function Network
26. rf: Random Forest
27. rpart: CART
28. rpart1SE: CART
29. rpart2: CART
30. RRFglobal: Regularized Random Forest
31. sda: Shrinkage Discriminant Analysis
32. sdwd: Sparse Distance Weighted Discrimination
33. slda: Stabilized Linear Discriminant Analysis
34. stepQDA: Quadratic Discriminant Analysis with Stepwise Feature Selection
35. treebag: Bagged CART
36. xyf: Self-Organizing Maps