Reviewer's report

Title: The effect of mode and context on survey results: Analysis of data from the Health Survey for England 2006 and the Boost Survey for London

Version: 1 Date: 18 June 2010

Reviewer: Jeanette Ziegenfuss

Reviewer's report:


This manuscript presents a comparison of results from two different survey protocols implemented in the same population with the same frame and method for drawing the sample and recruiting participants. The experimental design is a strength of this study. The question that the authors ask is an important one for survey methodologists and is of practical importance. The authors compare the survey results using two different strategies designed to account for differential nonresponse across the two protocols; nonresponse weights and propensity score matching. While I thought that the manuscript is an incremental contribution, my major concern was that there were many missing elements (i.e., what was the total sample size, what was the response rate for each protocol, for what characteristics did the nonresponse weights adjust). While I understand that there is a companion article, this article needs to stand on its own and as such should include these essential details. I also viewed the propensity score matching as a test of the robustness of the findings from the nonresponse weights. As such, I thought that these could be minimized and just referenced in that manner.

As interest in reducing costs was a major motivator for the selected design for the boost survey, I wonder if the authors could comment on the relative costs of the two methods.

Major compulsory revisions:

1. In the conclusion of the abstract you state that “…some data collected by interview and self-completion can safely be combined.” In the second paragraph of the strengths and limitations, you state that “…differences in survey estimates were not due to difference in sample composition or differential nonresponse but due to differences in measurement error.” While you have shown that this is the case for some variables under this particular protocol, it is possible that they will not hold in other scenarios. There are many things that could be driving potential differences in measurement across the two protocols, including mode effects and mode preference (i.e., differential nonresponse). Making nonresponse adjustments attenuates this in some cases, but not others. More discussion is needed around competing hypotheses. Isn’t it possible that there is still differential nonresponse on domains for which you do not address that result in
the differences you see for some variables or even artificially introduce the
sameness you see in other variables?

2. In the introduction, 1st paragraph, you state: “… the effect of survey mode on
comparability of health data is unknown.” There is a body of literature that speaks
to this issue, some of which you cite in your discussion. This literature should be
drawn upon here and the statement should be changed.

3. In the analysis, as stated above, more details are needed about the
nonresponse weights. This is an essential component of this manuscript and as
such it is important that the reader knows for which characteristics adjustments
were made.

Minor essential revisions:

1. In the second paragraph of the introduction you state that “after non-response
weights were applied, both samples provide reasonably close correspondence
with the London population for the characteristics examined…” While you never
state what characteristics you use to create your nonresponse weights, if they
are the same as those that you are comparing on, this is definitionally true.

2. It is not clear why the results are presented separately for men and women. 
Please defend this.

3. In the first paragraph of the methods, it is not clear if the nurse visit is at a
separate time than the initial visit. Do both visits have to occur for an interview to
be complete? Generally the design of the core and the boost – and which
elements are the same – is hard to follow. Might a figure help clarify?

4. In the second paragraph of the methods, you state that participants had the
option to return the questionnaire by post. As this would substantially lower the
costs of data collection as no second visit is needed, please comment on how
many did this.

5. In the results, the response rate and the total unweighted n by protocol should
be reported.

6. In the seventh paragraph of the results, you state that “the differences may
have been caused by the context in which the questionnaire was delivered.”
While this is true, it may also be due to differential nonresponse between the
core and the boost. While the nonresponse adjustments partially attenuate this
concern, it is possible that there are some differences that are not adjusted for
(such as underlying mode preference). This point should be discussed here and
throughout.

Discretionary revisions:

1. The use of Box 1 seems unnecessary. Much of the detail can be omitted and
the text could be integrated into the methods.

2. The last paragraph of the results describes item nonresponse. This was not
introduced earlier and seems more closely aligned with the companion article. I
would suggest omitting it from this manuscript.

3. I found the division of tables by those that were impacted and those that were
not odd. I think that it would make more sense to organize content into meaningful categories regardless of significance of findings.

Minor issues not for publication:
1. In the fifth paragraph of the discussion, insert “could” between “This” and “explains the apparent paradox…”
2. In the sixth paragraph of the discussion, “There was, however, no difference in the distribution of the responses in the core sample…” Is this supposed to be the boost sample? If not, I do not understand the conclusion as both groups in the core had another person present. Please change or give further explanation.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.