Author's response to reviews

Title: Application of Latent Semantic Analysis for Open-Ended Responses in a Large, Epidemiologic Study

Authors:

Travis D Leleu (travis.leleu@med.navy.mil)
Isabel G Jacobson (isabel.jacobson@med.navy.mil)
Cynthia LeardMann (cynthia.leardmann@med.navy.mil)
Besa Smith (besa.smith@med.navy.mil)
Peter W Foltz (pfoltz@pearsonkt.com)
Paul Amoroso (pjamoroso@aol.com)
Marcia Derr (marcia.derr@pearsonkt.com)
Margaret AK Ryan (margaret.ryan@med.navy.mil)
Tyler C Smith (tyler.c.smith@med.navy.mil)

Version: 2 Date: 8 July 2011

Author's response to reviews: see over
July 8, 2011

Adrian Aldcroft, BA  
Executive Editor  
BMC Medical Research Methodology

Dear Editor Aldcroft,

Enclosed please find our revised manuscript, “Application of Latent Semantic Analysis for Open-Ended Responses in a Large, Epidemiologic Study.”

The authors appreciate the opportunity to respond to the reviewers’ comments and all authors agree this revision has improved the clarity of objectives in the original manuscript. A revised version of the manuscript has been submitted. Our responses to each of the reviewers’ comments (in bold) follow below:

**Reviewer 1:**

1) **May be in the abstract you should mention the cluster analysis as I think it is the most relevant issue in this research (because you find 20 important areas of concern). I think Chi-square differences mention could be quit from the abstract.**

The authors agree the cluster analysis is more relevant than the chi-square analysis and should be included in the abstract. The text regarding chi-square has been removed and text regarding cluster analysis has been added to the methods section of the abstract.

2) **CLUSTER METHOD: Some questions and recommendations to validate the cluster solution and to see the reliability.**

   In the cluster analysis, why it is used a dissimilarity matrix (v.g., Euclidean distance) instead of a similarity matrix such as correlations or cosines matrix (more typical in LSA analysis)? To find clusters I think distance measures are very dependent on the vector length of the terms. Correlations and cosines are not dependent on the vector length. May be the clusters that emerge with similarity matrix (cosines) could be more relevant in medical terms.

   I think it would be a good idea to try with other method as Two step cluster and compare results to see if the 20 cluster solution is consistent as validate method.

We used a dissimilarity matrix, because the pam (partitioning around medoids) R function to create clusters requires a distance matrix, not a similarity matrix (there is a simple transform between the two, anyway). In particular we used $\sqrt{1 - \cos(v1,v2)}$ to create the distance matrix, where $v1$ and $v2$ are the LSA vectors of the two terms. So in fact the clusters are based on cosines among terms...
not on Euclidean distance. We have added clarifying information to the text of the first paragraph starting near the top of page 7.

3) Why it is chosen 20 clusters? Have you tried with other number of clusters?

We did try other numbers of clusters (10, 30, 40, 50). More than 20 gave redundant or overlapping clusters (e.g. clusters with medoids of ‘elbows’, ‘wrists’, ‘tendonitis’) or clusters that were not relevant to the medical domain (e.g. measures of time, military terms). Fewer than 20 clusters did not provide enough separation into relevant categories. We have clarified these points in the text also in the paragraph on page 7.

4) Finally, I think you can find some reliability measure between LSA and human expert. For example, you can choose ten randomly descriptions from each area of concern and give such descriptions to one expert. If the expert classifies the description in the same areas as LSA did, you will find a reliability measure that consolidates these areas (you can use Kappa coefficient).

We ran an additional analysis for human/LSA comparisons, asking human experts to categorize responses open-ended responses. We randomly selected 50 open-ended responses from the five most frequently reported areas of concern shown in Table 4 (illness/injury, exposure, exercise, back pain, and deployment) and asked two doctors on our staff to categorize these responses. One reviewer holds a PhD in epidemiology, and the other reviewer is preventive medicine physician (MD). Like the LSA technology, the reviewers were able to place each response in more than one category if they felt this was appropriate. The Kappa coefficients for the reviewers versus LSA were either substantial (0.6-0.8) or almost perfect (0.8-1.0) for four of the five areas of concern. The illness category had low agreement, both between the reviewers and LSA, as well as between the two reviewers. On the whole, the expert versus LSA reliability was excellent and provides further reassurance for using the LSA technology. Text was added to current page 14 in the paragraphs describing study strengths to briefly discuss this analysis.

5) What is exactly the hypothesis you have tested with Chi-square? I don’t understand which differences have you find.

The null hypothesis tested with the Chi-square is that there is no association between the two categorical variables. The authors agree that it is not clear which differences are presented in the manuscript. Our original chi-square analysis was to compare the differences by each characteristic shown in table 1 by response to the open-text field (yes vs. no). However, while the information for those not responding to the open-text field can be calculated from the numbers provided in Table 1, these numbers are not shown and therefore likely confusing to the reader. Since the results for the chi-square analysis are not mentioned in the results section for Table 1, and are also not mentioned in the discussion, the authors have removed the text regarding chi-square analysis from the abstract and the methods, statistical analysis section, near the top of page 8 to avoid confusion.
Reviewer 2:

1) Sorry if I missed it, but the foremost question this paper raised for me was whether the open responses were supplying new information or reiterating or reemphasizing “old” information. The authors correlate probability of response to self-reported health status and other covariates, but aren’t there many other relevant questions in the MCS?

The majority of responses provided in the open-ended text field were related to questions or topics already covered on the Millennium Cohort questionnaire. However, there were some responses to the open-ended questions that represented areas that were not covered on the survey, as highlighted in the discussion starting near the top of current page 11 such as some specific deployment-related exposures, disability, and injuries. We have since added questions on these areas to the survey. There were also a small proportion of comments expressed about very specific chemical exposures and outreach or government programs, which we do not directly address in our survey instrument mainly because they are beyond the scope of the instrument. We added text to current page 13 to add clarity stating “There was a very small proportion of responses related to very specific chemical exposures or other topics that were outside the scope of the survey, or very specific to a few individuals.” In general, the responses from the open text field were already covered on the survey in some form, but some more specific concerns that were not addressed were also provided. While there are many other relevant questions on the survey that could correlate with response, because this study was exploratory in nature, our aim was to understand the basic demographics and health status of participants responding. The authors feel that a more focused follow-up study would be a better platform for exploring additional potential correlates of response to the open-ended text field.

2) Do readers need both Tables 1 and 2, or would just Table 2 suffice? Of the two, I prefer to read the latter and felt the former was a little too long and unclear to be very helpful.

Table 1 provides a description of the population for the readers (e.g. how many men and women are in the total study population, and how does this compare with the proportion that answered the open ended text field?). Table 2 provides statistical measures of likelihood for which populations are more or less likely to answer the open-ended questions. The authors feel that without the information in Table 1, the reader may feel like they do not understand who comprises the population in the study. Therefore, we respectfully request to have Table 1 included in the manuscript, but defer to the editor for their opinion on this matter.

3) “Adjusted odds of response” on page 9 and in Table 2 mean that you’re running a single multivariate logit/logistic regression in each of the columns of Table 2, right? A clarifying note in the table or a sentence on page 9 might help.

We have added a sentence to the methods, statistical analysis on page 8 stating that a separate model was run for each population, and a sentence to footnote ‘a’ was also added to the table.

4) On page 9, I would highlight how education level does not seem to affect propensity to write-in, which is pretty interesting. This is apparently not a
“brainiac” or “blabber-mouth” effect.

We have added a sentence to the results on page 9 indicating this. We also updated the discussion section on page 12 to address the lack of association between education and response to the open-ended field.

5) Page 10, bottom: This tiny blurb about how (I think) folks filling out surveys in 2004-2006, when the Iraq War especially was really starting to go south, and concerns about deployment and mental health was fascinating to me. This kind of thing is probably what the authors are already working on next, and it’s what I had been hoping to see discussed here in much greater detail. I see it makes it back in on page 12, but basically as a reiteration.

Thank you for the suggestion. We agree that this is one of the next steps to take in this analysis. The granularity required from the LSA tool, make this idea very complex. We believe that by expanding the time period to incorporate future survey cycles will yield sufficient datasets that separating out subsets (e.g., deployed responders) will be more feasible. We have augmented the sentence at the top of current page 13 text to read “With only one follow-up data point available for the present study, it was difficult to fully understand this relationship; however, it will be interesting to examine whether these concerns persist at the same or increased levels in the 2007–2008 and future assessments”, to reaffirm that this concept would be of interest to the authors to examine with more follow-up data.

6) Table 2: I would add asterisks for statistical significance here, to help with clarity.

Thank you for this suggestion. Asterisks have been added to aid in locating statistically significant measures with ease.

Thank you for the opportunity to submit our paper for review. We appreciate your consideration.

Very Sincerely,

Travis Leleu
Senior Programmer
Deployment Health Research Department
Naval Health Research Center, San Diego, CA
Phone: 619-553-7317
Fax: 619-553-7601
Email: Isabel.jacobson@med.navy.mil