Preliminary Comments on the ISA from Dr. Lianne Sheppard

Comments on Chapter 3

- Overall the new organization represents good progress although I am not convinced that it meets its overall objective as outlined in Dr. Vandenberg's January letter. There are details and overall perspectives in the various sections that still need work and many of my comments address specific details. While this chapter is improved from the previous version, it is still a difficult chapter to digest. I suggest EPA take our suggestions for improvement, integrate them as best they can in the revised ISA, and more importantly, consider how to make further improvements in the ISA for the next criteria air pollutant to be reviewed. Ideally this chapter should become better with each new criteria pollutant ISA since a decent fraction of the material in this chapter applies to all criteria pollutants.
- P. 3-2 lines 20-27: This text needs to be improved. The definitions of differential and non-differential error and their link to misclassification need correcting.
    - It is too simplistic to think of differential error as systematic error. The technical definition means that the mismeasured covariate has information about the outcome beyond what is contained in the true exposure. Both differential and non-differential error are defined in terms of the information contained in the mismeasured exposure covariate after conditioning the outcome on both the true exposure and other covariates in the model not measured with error. With differential error information about the outcome remains; with non-differential error it does not. See Carroll et al, (2006, 2nd edition) *Measurement error in nonlinear models: a modern perspective*, Section 2.5 and Baker et al, (2008) *Environmental Epidemiology: study methods and application*, the Armstrong chapter on measurement error (chapter 5), section 5.2.
    - It is correct that exposure misclassification refers to categorical covariates while exposure measurement error refers to continuous covariates. Often, when speaking generally about the topic, the term exposure measurement error will be used for both. Differential and non-differential refer to both measurement error and misclassification.
- P 3-2 line 29: Some correction is needed. The previous sentence referred to the effect estimate (i.e. the estimate of the target parameter of interest in a disease model for an epi study), but this sentence is defining bias (and precision??) in terms of the exposure data.
- P. 3-2 line 33: The attenuation of effect estimates is characteristic of (pure) classical measurement error, which is a type of nondifferential error. Pure Berkson error doesn't tend to attenuate effect estimates (much, or at all – it doesn't bias the effect estimates at all when the disease model is linear).

- Pp 3-2 – 3-3, paragraph starting line 36: I appreciate the distinction being attempted regarding measurement error and a different target parameter of interest in the disease model. I'm not sure that I would call this "exposure error" and then go on to define classical and Berkson error. Also, I would drop the reference to the central site monitor measurement (p 3-3 line 7) as it depends on the context as to whether or not I would agree with this being an example of Berkson error.
- Section 3.2.3 gives useful perspective and is concise.
- Section 3.3.2 is still difficult to digest in terms of the goals of this document. I think we want to understand 1) what are different approaches to exposure modeling and 2) to appreciate how these various approaches affect the conclusions to be drawn from the epidemiological studies. The document has made some progress on 1) and could be better developed with respect to 2). (See also my comments on section 3.4.) However, regarding 1), improvements would still promote better understanding. I suggest dividing the discussion into approaches that focus on modeling measurements of SO2 vs those that focus on other ways of estimating ambient SO2 that may not use any measurement data (except perhaps to validate the modeling) but instead rely on physics, chemistry, emissions, and other information to predict concentration. Right now these are intermixed with LUR and IDW in the middle of SPM, dispersion models and CTMs. Regarding 2), I believe that the important idea is that the predictions from these models are used in epidemiological studies, predominantly cohort studies, as an exposure. So for the latter point it matters whether the model outputs are reasonable for the application. This could be affected by many factors such as data duration and representativeness (e.g. is a LUR based on data collected during one 2 week period useful at all for an epi cohort study focusing on long-term average exposure?) as well as the quality of the model results. There are important distinctions between models that use measurement data directly (i.e. statistical models such as LUR, IDW) and those that rely on other characteristics rather than measurements (i.e. SPM, CTM, dispersion models). Recently developed measurement error correction methods address statistical models.
- Section 3.3.2.2: The discussion on LUR should say early on that the focus is on a long-term average pollutant measure. LUR models capture spatial variability (as is said), so they are most useful in cohort studies where the focus is on long-term average pollutant exposure. Overall the discussion of the examples does not provide enough context to help readers appreciate how LUR model estimates affect inference in epidemiology studies, or really even to understand the examples themselves at face value. For instance, the averaging time of the measurements is not mentioned in the discussion of any of the examples.
- P 3-9 lines 19-21: LUR models are used to *predict* exposure (concentration) at unmeasured locations. I wouldn't characterize this as "increasing heterogeneity in the spatial resolution" but rather LUR models allow the ability to characterize more completely the spatial variation, by predicting at arbitrary locations. In reality many regression predictions

(e.g. from a LUR) will result in smoother surfaces than measurements. The wording suggests the opposite.

- P 3-9 sentence starting line 22: the apparently agnostic list of metrics for characterizing the results from LURs misses some important statistical ideas. The in-sample estimates of R2 or RMSE are estimates of training error, not generalization error and thus don't tell us what we want to know about how a predictive model performs. Cross-validation is one tool for estimating out of sample performance of a model. Estimates of RMSE can be either in sample our out of sample.
- P 3-9 l 30: It is not the samplers that affect the focus of LURs on long-term averages.
- P 3-10 l 7: I'm not sure how this example makes the intended point. From an epi point of view, better predictions don't necessarily give better health effect estimates. (see Szpiro et al 2011 Epidemiology) So citing an improvement in R2 doesn't necessarily help us understand whether or not LUR model results will be useful in epi studies.
- P 3-10 paragraph starting line 11: The Atari example discussion could be improved by: saying how the out of sample assessment was done, including how many monitors were used for that. Also since the sampling was for only one 2-week period, how does this reflect other times or a longer averaging period? How would these results be useful in an epi study?
- Pp 3-10 – 3-11: The discussion of Kanaroglou omits mentioning that the data are from a mobile platform and that the spatial autocorrelation was used to improve the model. Further, using the autocorrelation did not improve the predictive ability of the model very much. Also why is the focus of the discussion on statistically significant predictors?
- Section 3.3.2.3: Based on Table 3-1 this section covers kriging too. I see little mention of kriging and the term is not even defined. It should also be clarified whether the authors are referring to simple, ordinary, and/or universal kriging.
- Section 3.3.2.6. Even if researchers publish papers that contradict this point, it is not useful to apply stochastic population exposure models, or any models that simulate exposure, for epidemiologic inference. This is akin to doing multiple imputation without incorporating the outcome variable in the imputation. It leads to biased inference much like the effects of classical exposure measurement error. Stochastic population models are very useful for risk assessment.
- Section 3.3.3 is quite short and is mostly used to reference Table 3-1. Based on the heading, the purpose of this table is to describe metrics relevant only to epidemiology. This is reasonable. (However, this implies that stochastic simulation models should not be included, as is implied in coverage of the micro environmental models.) The new table represents an improvement in this version, though some work can be done to refine its content. I suggest that one way to improve this table is to recognize there are two components to exposure assessment for epidemiologic applications: 1) design and collection of exposure data and 2) using the data or deterministic models to produce exposure estimates for study subjects for use of epidemiologic inference, e.g. by using a prediction model. The first column, exposure

assignment method, lumps the source of data and the method for coming up with an estimated exposure for an individual in one category. The epidemiologic application column looks useful, but a bit notional. If there is no reference that can be cited to connect to an application being reviewed for SO2 epidemiology (in this or previous SOx ISAs), then I think that particular application should be omitted. The errors and uncertainties column appears to refer both to the exposure metric as well as to the impact on epidemiology. Finally, I suggest that Section 3.3.3 provide more of a road map to understanding each column in Table 3-1 particularly with respect to what criteria are used to fill in the cells of that table.

- Table 3-1: There are some questionable statements made in this table. For instance, that few input data are needed for IDW and kriging (presumably this is ordinary kriging, not universal kriging?) is misleading and if true almost certainly guarantees that the limitation noted (that it doesn't fully capture the spatial variation)
- Section 3.4: In principle this entire section starts off well, but it seems to get bogged down in detail that doesn't necessarily help readers understand the key issues for epidemiologic inference. Perhaps this is my misunderstanding because instead this section is intended to cover exposure both as it informs epidemiology AND as it informs risk assessment? This should be clearer to the reader. Also, in the context of epidemiology, by not distinguishing major study designs early, some sections are more confusing than necessary. For instance, the presence of spatial variation (section 3.4.2.2) makes using a single monitor in a time series study more problematic for SO2 than other pollutants (e.g. PM2.5) because of spatial variability. However, with sufficient data to model it, the spatial variation is an asset for SO2 in cohort studies. By not bringing study design in by now, the discussion of spatial variation is less helpful than it would otherwise be.
- Section 3.4.1.1: It would be helpful to provide some overall (perhaps closing) comments to help a reader use the discussion of AER in this section to better understand SO2 epidemiology.
- P 3-24 line 9: Is this statement true that ambient SO2 concentrations from central site monitors are commonly used as exposure surrogates in epi studies? I would agree for the time series design, but I need more evidence before I would agree that such a broad statement applies to all designs. I suggest references are needed to the studies reviewed in this document, or a less sweeping statement should replace this one.
- Section 3.4.2.1: A review of activity patterns seems particularly pertinent for risk and exposure assessment. I'd like to see the purpose of this section clarified. Is it for risk assessment, epidemiology, or both? If for epidemiology, what insights do we gain from the material in this section?
- P 3-27 lines 14-16: I suggest the sweeping statement about spatial variation leading to exposure error should be revised. In cohort studies we take advantage of spatial variation in long-term averages to make inference about health effects. Not capturing this well can lead

to Berkson(-like) or classical(-like) measurement error. But without leveraging this spatial variation we have no information with which to infer SO2 health effects in cohort studies.

- P 3-27 paragraph starting line 20, as well as much of the remainder of Section 3.4.2.2 on spatial variability: this discussion seems to be referring to time series studies without stating this.
- Section 3.4.2.3 Temporal variability: This is clearly focusing on the time series design. Thus the discussion is easier to follow and better tuned to the goals of the chapter.
- Section 3.4.3: The Zeger et al 2000 paper focuses on the time series design. The discussion should make this clear. The properties of the health effect estimate discussed should apply more generally to other designs, so I expect only a modest clarification is needed in this section.
- Section 3.4.3 Copollutant relationships: It will be important to understanding to recognize that correlation between measurements at the same location is only one aspect of understanding co-pollutant relationships, at least in time series studies. Pollutants that vary dramatically over space (as SO2) will have even lower correlation with other pollutants as the distance between locations increases, due to the inherent spatial heterogeneity of the pollutant. I'm not sure whether the full discussion in the subsections reflects this understanding.
- P 3-37 lines 19-20: I would only agree that SO2 exhibits a relatively high degree of exposure error compared to other criteria pollutants for the time series study setting that relies on one central site monitor, particularly in an area with SO2 point sources.
- P 3-37 paragraph starting line 24: I assume temporal copollutant correlations is referring to the correlation between pairs of pollutants measured at the same location, with repeat measurements over time used to estimate the correlation of these co-located pairs. I am finding it difficult to understand the sentence that starts on line 28: Is this for correlation of spatially varying pollutants, as one might have predictions of two pollutants at various locations in space? And what does "within-pollutant variation across space" mean here?
- P 3-38 lines 4-6: I assume these sentences describe correlations between pairs of measurements **at the same location**. Correct? The captions or footnotes for Figures 3-4 through 3-7 should mention the number of sites included in the analysis. (I recognize this may vary by pollutant, so perhaps this gets added to the side of the figure.)
- P 3-44 line 35: Please specify the within-daily time scale being referred to.
- P 3-45 lines 4-6: I think this sentence is conflating measured relationships with underlying relationships to draw conclusions about epidemiology results. What actually matters is how correlated the population exposure to SO2 is with other co-pollutants. The observed correlation at a single site may not reflect that, particularly in areas with a large amount of spatial heterogeneity of SO2.

- Figure 3-8: Do I understand correctly that each data point in each boxplot represents the correlation between pollutants as reported in the references listed on page 3-44? And was there any attempt to address whether or not monitors were co-located in this reporting of correlations, at least for the shorter-term measurements? And what do the data in the long-term correlations plot represent? How many studies are reporting in each row? (Do the dots reflect the raw data on these plots? Note: My copy of the ISA does not have any color on Figure 3-8, so it appears there are no red dots shown.)
- Figure 3-9: The data are approximate (because the median year for a study covering multiple years is used) and the scatter is large. I expect none of the regression lines plotted have slopes different from 0. I suggest one consider dropping this analysis, figure, and discussion as probably uninformative.
- Page 3-48 and Figure 3-8: I need help understanding the source of the correlations for the long-term estimates. Are these for long-term average predictions for pairs of pollutants at different participant residence locations? Or for co-located measurement pairs across monitors?
- Section 3.4.3.2 would benefit from a statement about the implications for SO2 health effect studies.
- P 3-49 line 26-7: What is a "surrogate target parameter of interest"?
- P 3-49 line 28: Pure Berkson error (U) is the unobserved part of the true exposure (T). The observed part is X. I.e. T=X+U. Berkson error is independent of X, and has mean 0. See the chapter by Armstrong in Baker et al 2008: *Environmental epidemiology*.
- P 3-49 line 30-1: Yes, for a linear disease model. There will be some specification bias in a nonlinear disease model. (I suggest just softening the statement to say something more like "Pure Berkson error generally does not bias...." There are some nuances with Berkson-like measurement error.)
- P 3-50 line 2: Replace "but" with "and".
- P 3-50 line 5: Insert "pure" before "Berkson"
- P 3-50 lines 25-27: I think the authors mean that nominal coverage of the CIs is below 95% for exposure effect estimates conditional on mismeasured covariates.
- Section 3.4.4.2 Long-term cohort studies: This section assumes that the ambient concentration value used in these studies is the concentration at the central site monitor. This may be true for older studies, such as the ACS and 6 cities studies, and even the more recent WHI study (Miller et al 2007). However, more recently cohort studies are using predicted exposures from a model, where ambient measurements come from multiple monitors. Further, the theoretical literature on measurement error impacts in cohort studies make the latter assumption that the exposure metric used in the epi study analysis is predicted from a statistical model.
- P 3-56 line 3: For clarity I suggest replacing "estimate" with "prediction"

- Section 3.5 Summary and conclusions:
    - Corrections to previous sections should be brought forward into this section
    - I still need to add some comments on the summary presented.