Discussion of “Using PIAAC Data for Producing Regional Estimates” by Kentaro Yamamoto

David Kaplan
Department of Educational Psychology
University of Wisconsin – Madison
Overview of Discussion

- Can the conditioning model be employed for small-area (regional) estimation?
- What is small-area estimation?
- Are there complimentary/alternative approaches?
Can the model be used for small-area estimation?

- Tentatively … yes!
  - One needs a large amount of background information and preferably a large small-area.
  - This is true if there are few responses to the cognitive items.
  - Additional information coming from the cognitive item responses improves the situation.
What is small area estimation?

- Following Longford (2005) and Pfeffermann (2013)
  - The problem of small area estimation (SAE) is how to produce reliable estimates of parameters of interest such as means, counts, quantiles, for sub-populations for which only small samples or no samples were obtained.
What is SAE? (cont’d)

- The importance of SAE is due to the fact that such problems as education policy planning rely heavily on estimates from small areas, even though these small areas were not part of a large sampling scheme.
- It may simply have been too expensive to sample all small areas of interest.
What is SAE? (cont’d)

• Note that small-areas are not necessarily geographic.
  - Sample size constitutes the problem, not geographic size.
  - “Areas” are just sub-populations – could be socio-demographic groups.
What is SAE? (cont’d)

- Two types of SAE approaches
  - Design-based
  - Model-based
- Both require the use of auxiliary information from large surveys.
- Kentaro’s paper offers a model-based SAE method, based on the conditioning model.
What is SAE? (cont’d)

- SAE exploits “similarity” among the sub-populations of the country.
- SAE “borrows strength” from other sub-populations via a hierarchical model.
- A model is estimated on similar sub-populations using auxiliary information available for all subpopulations.
- SAE yields “synthetic estimation” for non-sampled sub-populations.
What is SAE? (cont’d)

• Notice that this is essentially a missing data problem.
• We are missing sufficient (or any) data for precise inferences.
• Creating synthetic data for LSAs has been tried (Kaplan & McCarty, 2013) with the OECD PISA and TALIS surveys.
• Kentaro’s paper is the first to try SAE using the conditioning model.
Kentaro applies the conditioning model for proficiency estimation applied to a sub-population (a “small-area”) – namely Ontario.

His method builds on the similarity of covariances among exogenous and endogenous variables between the national population and sub-population.

We are not provided with a method for how similarity is assessed, or if similarity was obtained in this example. Extant literature gives little advice on assessing similarity.
What does it mean to “assess similarity”?

- One way to see the similarity issue is through the lens of a missing data problem.
- We are interested in estimating proficiency scores for individuals in a sub-population based on using a conditioning model that was calibrated and estimated on the national sample.
- Missing data approaches that formally assess similarity might be useful.
Statistical matching

• Essence of the matching methods
• Find a member of the sub-population who is “close” to a comparable member of another sub-population or national population on background characteristics and (perhaps) item responses.
• Provide the sub-population member with the proficiency values obtained from the matched target member. This yields a small-area synthetic file.
One Approach: Hot Deck Matching

- The hot deck algorithm finds a person in the sub-population that is closest to a person in the national population in terms of background questionnaire data and cognitive responses.
- “Close” could be assessed using a multivariate Euclidian distance.
- If more than one person is found – choose one at random.
- The idea is to assess similarity by assessing the number of matches.
Similarity Measures

- Euclidean Distance
- Maximum distance
- Mahalanobis distance
- Predictive Distance
Matching methods

• Some examples available in R computing environment
  - Predictive mean matching (mice)
  - Bayesian bootstrap predictive mean matching (BaBooN)
  - EM – Bayes bootstrap hybrid (Amelia)
Future research

- Multiple model-based methodologies are available for small-area estimation.
- Extensions of Kentaro’s paper would include:
  1. Assessing similarity in the context of the conditioning model
  2. Comparing a variety of similarity measures and matching approaches
Future research (cont’d)

• The question is how to employ the conditioning model and assess similarity.
  - Are all background and cognitive responses used, or just the background conditioning variables?
  - How should plausible values be handled?
Future research (cont’d)

- Can an approach to assessing the validity of the procedure be developed?
- Validity tests exist for data fusion (Rässler, 2002; Kaplan & McCarty, 2013).
- Are these tests appropriate for employing the conditioning model to a small-area?
- One case study is definitely not enough.
Conclusion

• There is always interest in extending the utility of LSAs.
• There is a great deal of basic research that needs to be supported and conducted.
• Kentaro’s paper is an important first step in the work on small-area estimation in the context of proficiency estimation in large-scale assessments.
Thank you for your attention

Richard_Murnane@Harvard.edu