Peer Review of Highly Influential Scientific Assessments

An Information Quality Bulletin of the Office of Management and Budget, dated December 16, 2004, and published in the Federal Register on January 14, 2005 (pages 2664–2677), required federal agencies to conduct a "peer review" of "influential and highly influential scientific information," as those terms are defined in the Bulletin, prior to dissemination to the public.

The *Housing Choice Voucher Program Administrative Fee Study* sponsored by the U.S. Department of Housing and Urban Development, Office of Policy Development and Research (PD&R), may constitute "influential scientific information," under the terms of the Bulletin.

In accordance with the previously described guidelines, PD&R asked two industrial engineers who are experts in time-and-motion research (Dr. Nicola Shaw and Dr. Kai Zheng) and one economist who is an expert in assisted housing (Dr. Edgar Olsen), to review a draft of the *Housing Choice Voucher Program Administrative Fee Study Final Report*.

PD&R asked the three peer reviewers to keep the following questions in mind when reviewing the study:

1. Does the study respect the norms of the profession in collecting and analyzing data?
2. Does the study respect the norms of the profession in testing theory against evidence?

After assessing the initial peer review comments received from each of the three expert reviewers, the Abt Associates Inc. team made edits to the Final Report and drafted formal responses to comments from Dr. Nicola Shaw and Dr. Kai Zheng. PD&R sent the Final Report, the formal responses to comments, and a response to two particular comments made by Dr. Edgar Olsen, to the three expert reviewers. The expert reviewers then sent their final comments to PD&R. The initial comments, responses to those comments, and final comments for each of the three peer reviewers appear on the following pages.

Initial Peer Review Comments from Dr. Nicola Shaw, 15 May 2015:

Thank you for asking me to act as a peer reviewer for the **Housing Choice Voucher Program Administrative Fee Study: Draft Final Report**. Specifically, I was asked to consider two questions focusing my attention on the investigators' use of Random Moment Sampling (RMS):
1) Does the study respect the norms of the profession in collecting and analyzing data?
2) Does the study respect the norms of the profession in testing theory against evidence?

Unfortunately I can't answer these questions, at this time, as 1) the draft report doesn't provide enough information and 2) the second question isn't an appropriate one for this specific study at this point in time.

In terms of the first question; usual statistical sampling methods for RMS require that the sample should cover the entire time period involved in the sampling universe. Page 26 of the Draft report states that *"The sample frame was the employee's day-to-day work schedule over the 40-working-day interval for data collection. RMS notifications were drawn randomly within 36 minute-blocks from this defined frame…"* Consequently, I don't know what is meant by 36 minute blocks. Do the investigators mean 36 different time periods of a minute each, or a number of time periods of 36 minutes duration each. Therefore, I can't tell if all possible moments had an equal chance of being selected. Further clarification is needed.

Likewise, Page 26 states that *"Staff were notified 12 to 15 times a day (about 1.5 notifications per hour)…";* however, no explanation of why this number of data points was selected is provided. Again, this means that it is unclear whether or not all possible time points had an equal chance of being selected.

Statistical sampling methods require that invalid responses be dealt with appropriately. Page 34 of the Draft report states that *"Any after-hours notifications that were not answered within four hours would disappear from the device, unlike notifications during regular working hours, which would all need to be answered."* This suggests that invalid responses were ignored, rather than addressed statistically. As a result of these exclusions the sample results could be biased. This needs to be addressed.

RMS is usually conducted over a fiscal quarter as opposed to eight weeks. This helps to mitigate for time of year related differences in work activity. It is assumed that the use of a 40 working day data collection period was for fiscal reasons. This should be clarified and the potential impact of this recognised.

Page 35 states that *"Collecting time data from different PHAs at different times of the year allowed the study to measure program times and costs at different points in the program cycle, which is very important for ensuring that activities that do not happen very often or happen only once a year are not missed."* While true, the study protocol didn't take into account the fact that those activities which don't happen often, or only once a year, may occur at different points in the year in different PHAs so they may still have been missed through the approach used. The Investigators do recognise some of the problems with this approach by stating that *"The* *disadvantage of this approach is a higher level of variation across PHAs in time observed for different activities."*

In terms of the second question; this question is not truly answerable for this specific study, at this time. The Draft report presents the methods used to investigate the administrative costs of the HCV program and proposes an alternate administrative fee structure based on the results of the investigation undertaken.

As such, there is little evidence that theory could be tested against at this time. Evidence will only exist if the proposed fee structure is applied and data collected to demonstrate that the theoretical understanding underpinning the revised structure actually provided for a more appropriate distribution of resources.

One could argue that the comparison of the proposed fee structure to the existing fee structure (Page 147) is testing theory against evidence and, if so, this work certainly meets the accepted norms in this area. However, it should be remembered that while evidence can prove a theory wrong, it cannot necessarily prove a theory correct because there may be evidence, yet to be discovered, that exists that is inconsistent with the theory.[1]

Thank you for the opportunity to comment on this draft report. I look forward to receiving a copy of the final report in due course.

Dr. Nicola (Nikki) Shaw, FBCS, CITP.
ESRI Canada Health Informatics Research Chair: Health Informatics Institute & Algoma University
Associate Professor: Algoma University & Northern Ontario School of Medicine
Clinical Professor: Department of Clinical Sciences, School of Medicine, Caribbean Medical University (Curacao)
Adjunct Associate Professor: University of Ontario Institute of Technology, Lakehead University, Laurentian University, University of Victoria,
University of Alberta, University of British Columbia

<u>Responses to Dr. Nicola Shaw's Peer Review Comments, 9 June 2015</u>:

Comment 1:
In terms of the first question; usual statistical sampling methods for RMS require that the sample should cover the entire time period involved in the sampling universe. Page 26 of the Draft report states that *"The sample frame was the employee's day-to-day work schedule over the 40-working-day interval for data collection. RMS notifications were drawn randomly within 36 minute-blocks from this defined frame…"* Consequently, I don't know what is meant by 36 minute blocks. Do the investigators mean 36 different time periods of a minute each, or a number of time periods of 36 minutes duration each. Therefore, I can't tell if all possible moments had an equal chance of being selected. Further clarification is needed.

[1] K. R. Popper (1902-1994)

Response to Comment 1:
The revised report will clarify that 36 minute grids were used to segment the sample frame, and then random moments were drawn within each 36 minute grid. Every drawn moment was drawn with an equal probability of being selected.
Additions to the report are shown in italics:
Rather than using pure simple random sampling, the team employed gridded simple random sampling to ensure that each period of the day and each day of the week had sufficient sampling coverage. This design was used to address that HCV activities are not uniformly performed across different times of day or days of the week. To ensure that all working time was sampled with equal probability, the study team wrote a custom sampling tool that first stratified the entire eight week work schedule into 36-minute grids and then drew one and only one time randomly from within each grid. As in simple random sampling, each working time segment has equal chance of selection, but the selections are more uniformly distributed across each working day and each day of week with this stratification. The draws were made in advance on the server for each participant's work schedule and populated in a database then synced to the participant's device.

This systematic surveying of activities produced a count of notifications assigned to mutually exclusive functions and the total estimated time staff worked during regularly scheduled hours. The notifications were turned into total minutes of activity performed using time expansion with sampling weights. Since each RMS survey was drawn with equal probability within the 36-minute grids, each response represents 36 minutes of the sample frame (sample weight = 36 minutes). For a person who is scheduled to work 9 hours a day for 40 days, this represents 360 hours in the sample frame. With the 36-minute grids, the RMS method would have asked them approximately 600 RMS surveys in that period. If five of the resulting responses were "Meetings" for example, time expansion turns this into an estimate by multiplying 36 x 5 = 180 minutes.[2]

Comment 2:
Likewise, Page 26 states that *"Staff were notified 12 to 15 times a day (about 1.5 notifications per hour)…,"* however, no explanation of why this number of data points was selected is provided. Again, this means that it is unclear whether or not all possible time points had an equal chance of being selected.

Response to Comment 2:
The revised report will clarify that we used Power Analysis to achieve a suitable level of precision. The variation in moments drawn per day was directly dependent on the person's work schedule. Again, time was equally sampled for all staff for all days.
Additions to the report are shown in italics:
RMS designs aim to capture the variation in work performed in an appropriate period of time. This defined universe of employee time to sample is known as the sample frame. For this study, the sample frame for RMS was the employee's day-to-day work schedule over the 40 work-day

[2] In most cases, this is algebraically identical to proportional scaling, where 5 of 600 RMS notifications = 0.83 percent. Multiplying the total sample frame of 360 hours x 0.83 percent also yields 180 minutes.

interval for data collection.[3] *RMS notifications were drawn randomly within 36-minute gridded blocks from this defined frame, resulting in more than 600 notifications for a typical full-time PHA worker over 40 days. The rate of notifications was designed to detect small agency-level effects in time allocation and establish a high level of precision in time estimates using power analysis with the arcsine transformation for differences in proportions (Cohen 1988). The team assumed an acceptable Type I error rate of 5 percent and aimed for statistical power of 90 percent. The rate determined how many notifications were needed in total and subsequently each day to meet our statistical criteria.*

Comment 3:
Statistical sampling methods require that invalid responses be dealt with appropriately. Page 34 of the Draft report states that *"Any after-hours notifications that were not answered within four hours would disappear from the device, unlike notifications during regular working hours, which would all need to be answered."* This suggests that invalid responses were ignored, rather than addressed statistically. As a result of these exclusions the sample results could be biased. This needs to be addressed.

Response to Comment 3:
The revised report will include the following clarification about after hours work:
The reported work schedule of staff reflected the core RMS sampling approach. However, the team desired a data collection mechanism so staff could report any "after-hours" working time using the same smartphone tool. On rare occasions, PHA staff work outside of their regular working hours—coming in early, staying late, or working on the weekend. Although the prevalence of non-scheduled HCV work was known to be small, the team created an additional smartphone feature to issue RMS notifications into the evening and on weekends and activated this feature for participating staff who indicated that they sometimes work in these other periods.[4]

With the after-hours functionality, the study team was able to collect supplemental information to describe all work done by each staff member, even if that work occurred outside of normal working hours. However, this additional captured work effort accounted for less than one

[3] The study team considered longer RMS intervals including both 12 continuous weeks and 12 weeks broken into two waves at each PHA at different times in the year. Feedback from the EITRG in an early study design meeting suggested that extending data collection would be overly burdensome on the participating PHA staff, so there was little support for extending the approach. Dividing the data collection into two waves was believed be the strongest approach methodically but impractical to execute with the resources available. Also, PHA employees in the pretest indicated a strong opposition to participating for more than eight weeks.

[4] Staff received the "after hours" functionality if they indicated that they worked outside of their usual schedule at least once a week or four times a month. Staff were instructed to respond to the after-hours notifications only if they were actually working when they received one. Since the after-hour notifications canvassed a very large block of evening and weekend time, the team anticipated PHA staff would be answering "not working" the vast majority of the time. Thus, to alleviate burden, the team instructed PHA staff to ignore the notifications unless they were actively working. To further reduce burden, these unanswered responses were cleared from the device's calendar before they arrived to work the next day. The treatment of non-response in this after-hour window was interpreted as "not working" and coded accordingly during analysis. Brief follow-up interviews with PHA staff were completed to validate the extent of this after-hours work.

percent of all work time reported when combining with the RMS sampling done during regular work hours.

Comment 4:
RMS is usually conducted over a fiscal quarter as opposed to eight weeks. This helps to mitigate for time of year related differences in work activity. It is assumed that the use of a 40 working day data collection period was for fiscal reasons. This should be clarified and the potential impact of this recognised.

Response to Comment 4:
We added the following clarification about why we collected data over a two-month period rather than one fiscal quarter or some other timeframe:
The study team considered longer RMS intervals including both 12 continuous weeks and 12 weeks broken into two waves at each PHA at different times in the year. Feedback from the EITRG in an early study design meeting suggested that extending data collection would be overly burdensome on the participating PHA staff, so there was little support for extending the approach. Dividing the data collection into two waves was believed be the strongest approach methodically but impractical to execute with the resources available. Also, PHA employees in the pretest indicated a strong opposition to participating for more than eight weeks.

Final Peer Review Comments from Dr. Nicola Shaw, 10 June 2015:

Thank you for asking me to act as a peer reviewer for the **Housing Choice Voucher Program Administrative Fee Study: Final Report**. Specifically, I was asked to consider two questions focusing my attention on the investigators' use of Random Moment Sampling (RMS):
1) Does the study respect the norms of the profession in collecting and analyzing data?
2) Does the study respect the norms of the profession in testing theory against evidence?

The revisions made since the Draft version of this report have adequately addressed my concerns and clarified areas of potential confusion. Therefore, I am comfortable in now stating that this study does respect the norms of the profession in collecting and analysing data.

As with the draft report, the final version of the report leaves the second question not truly answerable for this specific study, at this time. The report presents the methods used to investigate the administrative costs of the HCV program and proposes an alternate administrative fee structure based on the results of the investigation undertaken.

As such, there is little evidence that theory could be tested against at this time. Evidence will only exist if the proposed fee structure is applied and data collected to demonstrate that the theoretical understanding underpinning the revised structure actually provided for a more appropriate distribution of resources.

One could argue that the comparison of the proposed fee structure to the existing fee structure is testing theory, against evidence and, if so, this work certainly meets the accepted norms in this area. However, it should be remembered that while evidence can prove a theory wrong, it cannot necessarily prove a theory correct because there may be evidence, yet to be discovered, that exists that is inconsistent with the theory.[5]

Thank you for the opportunity to comment on this report.

Initial Peer Review Comments from Dr. Kai Zheng, 18 May 2015:

The draft report, entitled “Housing Choice Voucher Program (HCV) Administrative Fee Study,” describes the design, results, and recommendations of the HCV Administrative Fee Study. The study was funded by the U.S. Department of Housing and Urban Development (HUD). The purpose was (1) to develop scientific evidence on the costs of operating a high-performing and efficient Housing Choice Voucher program, and (2) to devise a new administrative fee formula based on this evidence.

The study used a random moment sampling approach to quantify the time that staff of public housing agencies (PHAs) spent on performing core HCV program functions. A total of 909 PHA staff, from 60 high-performing and efficient PHAs across the country, participated by responding to random electronic prompts delivered via a smartphone device over a period of eight weeks. The study then used this information to estimate the costs of operating a high-performing and efficient HCV program. The costs considered in the study included labor costs, frontline non-labor costs, and overhead costs, and were adjusted for cost-cutting actions that might be a result of administrative fee proration and sequestration. The study further recruited 130 PHAs with fewer than 250 vouchers to participate in small program interviews to investigate whether there is a minimum number of vouchers below which a PHA cannot operate the HCV program on administrative fees alone.

Overall, the study was well designed and well executed. The report was well written. The focus on high-performing and efficient HCVs is appropriate given the study’s designed objectives. The rationale for choosing random moment sampling, over other available time measurement methods, was well articulated. The cost estimation and adjustments also appear to be sound. While the study made numerous assumptions and methodological compromises, and was constrained by the limitation of the data, the research team did a good job in ensuring the successful conduct of the study and robustness of the results.

Below are some high-level concerns regarding the study, some of which may be clarification questions.

First, it is not clear when and how the SEMAP high performance score criteria were applied. Were all 346 PHAs selected as “primary picks” or “backups” high performers according to the SEMAP criteria? The report does not explicitly state so. If not, the reviewer wonders why SEMAP scores were not used to pre-screen PHAs prior to soliciting their interest in participating, given that these scores are readily available?

[5] K. R. Popper (1902-1994).

Second, the report suggests "PHAs that did not meet the SEMAP high performance score criteria but that were determined to be high-performing HCV programs by HUD headquarters and field staff and recommended for inclusion in the study were also included in the sampling frame." According to Chapter 2 of the report, the 346 initially selected PHAs were first reduced to 297 after HUD review (49 PHAs were deemed as not qualified), and then reduced to 99 sites willing to participate, 95 sites visited, and 60 sites eventually included. It is not clear when the additional sites—those that did not meet the SEMAP criteria but were included in the study based on the recommendation of HUD headquarters and field staff—were introduced. If not all 346 PHAs selected in the initial phase met the SEMAP criteria, then when were the SEMAP criteria actually applied?

Third, the middle column in Exhibit 2.7 suggests that there were 1,258 high-performing HCV programs. How were these "high-performing" programs determined, especially given that the SEMAP criteria alone seem to be inadequate for the purpose of the study according to the report?

Fourth, the research team made some minor changes to improve the categorization of activities based on the pilot testing results at the four pilot sites. The reviewer wonders how these changes might affect the use of the data collected from these pilot sites?

Fifth, the report mentions that the notifications were turned into total minutes of activity performed using time expansion with sampling weights. It is not clear what time expansion procedure(s) were used, and how sampling weights were applied. Further, the report should provide more detail regarding how sampling was performed for larger PHAs where not all staff members were invited to participate (e.g., number of staff participated versus size of the staff).

Lastly, elaboration on the following concepts is recommended:

Moving to Work (MTW) demonstration project.

SEMAP high performance score criteria. Also, the acronym SEMAP should be defined prior to use.

Kai Zheng PhD
Associate Professor, Schools of Public Health and Information
Interim Director, The Health Informatics Program
University of Michigan

Responses to Dr. Kai Zheng's Peer Review Comments, 9 June 2015:

Comment 1:
First, it is not clear when and how the SEMAP high performance score criteria were applied. Were all 346 PHAs selected as "primary picks" or "backups" high performers according to the SEMAP criteria? The report does not explicitly state so. If not, the reviewer wonders why SEMAP scores were not used to pre-screen PHAs prior to soliciting their interest in participating, given that these scores are readily available?

Response to Comment 1:
The revised report will include the following clarification:
All 346 PHAs selected as primary picks or backups either met the SEMAP criteria (i.e., scored as high performers in three of the past four years or in at least two of the previous four years for those PHAs not rated each year) or were recommended by HUD headquarters or field staff.

Comment 2:
Second, the report suggests "PHAs that did not meet the SEMAP high performance score criteria but that were determined to be high-performing HCV programs by HUD headquarters and field staff and recommended for inclusion in the study were also included in the sampling frame." According to Chapter 2 of the report, the 346 initially selected PHAs were first reduced to 297 after HUD review (49 PHAs were deemed as not qualified), and then reduced to 99 sites willing to participate, 95 sites visited, and 60 sites eventually included. It is not clear when the additional sites—those that did not meet the SEMAP criteria but were included in the study based on the recommendation of HUD headquarters and field staff—were introduced. If not all 346 PHAs selected in the initial phase met the SEMAP criteria, then when were the SEMAP criteria actually applied?

Response to Comment 2:
The initial sampling universe included those PHAs that met the SEMAP criteria and those PHAs recommended by HUD. The revised report will include the following clarifications:
The PHAs had to meet one of two criteria to be included in the sampling universe:

- ***High Performance on SEMAP**: PHAs had to have scored as "high performers" on SEMAP for three of the past four years (at the time of sample selection) or in at least two of the previous four years for those PHAs not rated each year.*
- ***Recommendation by HUD Headquarters or Field Staff:** HUD headquarters and field staff were given the opportunity to recommend PHAs that they thought would be suitable for the study because they administered high performing and efficient programs.*[6]

Comment 3:
Third, the middle column in Exhibit 2.7 suggests that there were 1,258 high-performing HCV programs. How were these "high-performing" programs determined, especially given that the SEMAP criteria alone seem to be inadequate for the purpose of the study according to the report?

Response to Comment 3:
The middle column is the sampling universe. In the revised report, we will relabel the column as "Sampling Universe for the Study" and add the following footnote to the text immediately before the exhibit to remind readers what the sampling universe is:
As described in Section 2.1.1., the sampling universe is all non-MTW PHAs with HCV programs of more than 100 vouchers and several years of high performer ratings on SEMAP, plus HUD recommendations.

[6] Over the three rounds of sampling, HUD headquarters and field staff recommended a total of 119 PHAs, 29 of which did not meet the SEMAP criteria (i.e., high performer rating for three of the past four years or at least two of the previous four years for those PHAs not rated each year). Two of the PHAs in the final sample of 60 were HUD-recommended PHAs that did not meet the SEMAP criteria at the time of sample selection.

Comment 4:
Fourth, the research team made some minor changes to improve the categorization of activities based on the pilot testing results at the four pilot sites. The reviewer wonders how these changes might affect the use of the data collected from these pilot sites?

Response to Comment 4:
The changes were predominantly used to add more detail for PHA staff in the activity survey and did not impact the higher level categorization after the pretest sites to a meaningful degree. The vast majority of data analyzed was done so at higher levels of reporting, above where most changes were made so we don't believe anything impacted the data quality and equivalence for the study's main objectives. In a few instances, small wording changes were made to categories to ensure clarity.
Text added to the report:
The changes to the categorization generally involved adding more detailed categories of work activities. This made the data collection easier for PHA staff because they were able to recognize the components of their work more readily when they were not embedded in a broader category. However, in analyzing the time data we generally aggregated up to the original broader categories. Thus, the changes made following the pretest improved the experience for PHA staff but did not affect the analysis.

Comment 5:
Fifth, the report mentions that the notifications were turned into total minutes of activity performed using time expansion with sampling weights. It is no clear what time expansion procedure(s) were used, and how sampling weights were applied.

Response to Comment 5:
The revised report will include the following clarification:
Rather than using pure simple random sampling, the team employed gridded simple random sampling to ensure that each period of the day and each day of the week had sufficient sampling coverage. This design was used to address that HCV activities are not uniformly performed across different times of day or days of the week. To ensure that all working time was sampled with equal probability, the study team wrote a custom sampling tool that first stratified the entire eight week work schedule into 36-minute grids and then drew one and only one time randomly from within each grid. As in simple random sampling, each working time segment has equal chance of selection, but the selections are more uniformly distributed across each working day and each day of week with this stratification. The draws were made in advance on the server for each participant's work schedule and populated in a database then synced to the participant's device.

This systematic surveying of activities produced a count of notifications assigned to mutually exclusive functions and the total estimated time staff worked during regularly scheduled hours. The notifications were turned into total minutes of activity performed using time expansion with sampling weights. Since each RMS survey was drawn with equal probability within the 36-minute grids, each response represents 36 minutes of the sample frame (sample weight = 36 minutes). For a person who is scheduled to work 9 hours a day for 40 days, this represents 360 hours in the sample frame. With the 36-minute grids, the RMS method would have asked them approximately 600 RMS surveys in that period. If five of the resulting responses were

"Meetings" for example, time expansion turns this into an estimate by multiplying 36 x 5 = 180 minutes.[7]

Comment 6:

Further, the report should provide more detail regarding how sampling was performed for larger PHAs where not all staff members were invited to participate (e.g., number of staff participated versus size of the staff).

Response to Comment 6:

The revised report will include the following clarification:
We included in the RMS data collection all the staff who served unique functions and sampled only among groups of staff who served the same function. The sampling approach used was a stratified simple random sampling of staff. The sampling strata were defined by staff role (inspectors, housing specialists, intake specialists, etc.). The study team worked with the PHA to identify each staff that played a unique role in the program and to ensure that all of those staff participated in RMS. For staff that served the same role in the program, we selected a random sample of those staff, with the number of staff sampled determined by the total number of staff that would participate in RMS (up to 100) and by the proportion of staff serving each role. For example, if there were 10 inspectors that played the same role and 30 housing specialists that played the same role, and we had enough resources to have 60 percent of the staff participate, we would randomly select 6 inspectors and 18 housing specialists to participate.

Comment 7:

Lastly, elaboration on the following concepts is recommended: Moving to Work (MTW) demonstration project; SEMAP high performance score criteria. Also, the acronym SEMAP should be defined prior to use.

Response to Comment 7:

The revised report will include the following clarification on MTW:
Moving to Work (MTW) is a demonstration project, enacted by Congress in 1996, under which a limited number of public housing authorities test ways to increase the cost effectiveness of federal housing programs, to increase housing choices for low-income families, and to encourage greater economic self-sufficiency of assisted housing residents.
The revised report will include the following clarification on SEMAP:
The first stage in selecting the sample was to identify the sampling universe based on ratings from HUD's Section 8 Management Assessment Program (SEMAP). SEMAP measures the performance of PHAs that administer the Housing Choice Voucher program across 14

[7] In most cases, this is algebraically identical to proportional scaling, where 5 of 600 RMS notifications = 0.83 percent. Multiplying the total sample frame of 360 hours x 0.83 percent also yields 180 minutes.

indicators.[8] *HUD assigns each PHA a rating on each of the 14 indicators and an overall performance rating of high, standard, or troubled. The ratings for indicators 1-8 are based directly on the PHA's certification to HUD. The ratings for indicators 9-14 are based on HUD administrative data. Metropolitan PHAs are able to earn bonus points for their achievements in encouraging assisted families to choose housing in low poverty areas.*

Final Peer Review Comments from Dr. Kai Zheng, 16 June 2015:

I looked over the responses and how the authors addressed the questions I raised. I think they did a good job in responding. I have no further concerns.

Initial Peer Review Comments from Dr. Edgar Olsen, 19 May 2015:

The primary purpose of the Housing Choice Voucher Program Administrative Fee Study is to develop a formula for allocating the fees for administering the Voucher Program to the local, regional, and state housing agencies that operate it. In the end, the study produces a formula by estimating a regression model that explains differences in the administrative cost per voucher across housing agencies with a similar level of performance in terms of factors beyond the agency's control. In assessing the study, I was asked to address two questions:

1) Does the study respect the norms of the profession in collecting and analyzing data?
2) Does the study respect the norms of the profession in testing theory against evidence?

In my opinion, it does both.

First, it reflects careful thought about factors beyond the housing agency's control that would affect the cost of administering the Housing Choice Voucher (HVC) program at a given quality level. The Abt staff involved in the study have considerable knowledge about the program's operation and the evidence on its performance, and they consulted with a large, knowledgeable advisory committee and many local officials who operate the program. In the end, they identified 58 potential variables in 7 categories that might affect administrative cost.

[8] The indicators are: (1) proper selection of applicants from the housing choice voucher waiting list; (2) sound determination of reasonable rent for each unit leased; (3) establishment of payment standards within the required range of the HUD fair market rent; (4) accurate verification of family income; (5) timely annual reexaminations of family income; (6) correct calculation of the tenant share of the rent and the housing assistance payment; (7) maintenance of a current schedule of allowances for tenant utility costs; (8) ensure units comply with the housing quality standards before families enter into leases and PHAs enter into housing assistance contracts; (9) timely annual housing quality inspections; (10) performing of quality control inspections to ensure housing quality; (11) ensure that landlords and tenants promptly correct housing quality deficiencies; (12) ensure that all available housing choice vouchers are used; (13) expand housing choice outside areas of poverty or minority concentration; and (14) enroll families in the family self-sufficiency (FSS) program as required and help FSS families achieve increases in employment income. SEMAP regulations can be found at 24 CFR 985.

Second, the empirical results are based on data for a random sample of housing agencies that had a roughly similar level of performance. The large cost of assembling data on the amount of time and money devoted to various administrative activities by a housing agency precluded including all housing agencies characterized as high performers in the study. In my judgment, the size of the sample (60 agencies) proved adequate to estimate the effects of the most important factors on the administrative cost of operating the HVC program at a given quality level.

Third, the study meticulously collected data for each housing agency in the sample concerning the amount of time devoted to various administrative activities and their cost and the factors that were hypothesized to affect administrative cost.

Fourth, the study used reasonable, systematic methods for selecting the subset of variables to be included in the final regression model whose estimation leads to the proposed formula for allocating the fees for administering the HCV program to the agencies that operate it. Because the study team identified 58 potential factors explaining differences in administrative cost per voucher and the sample contains only 60 observations, it is not possible to estimate with any precision the parameters of a regression model that includes all 58 factors. The study proceeded in steps to identify the most important factors and obtain credible estimates of their effects on administrative cost based on a combination of reasonable theoretical presumptions about the direction of the effect of each factor and statistical criteria, namely, the statistical certainty that the factor had some effect on administrative cost and its contribution to the fit of the regression model.

All regression models included two variables with a strong theoretical basis, namely, an index of the wage rate of government workers across locations and a measure of the size of the program. The statistical results consistently supported their inclusion.

To reduce the number of determinants of administrative cost to a reasonable number for further exploration, the analysts added the remaining potential explanatory variables to the regression model one at a time and chose variables for further exploration based on the degree of statistical certainty that they have an effect on administrative cost per unit beyond the effect of the local wage rate and the dummy variable for a small voucher program. This exercise pruned the list of factors to 20. Three were removed from further consideration for persuasive theoretical and statistical reasons.

Additional variables were removed because they were highly correlated with other variables that were different measures of the same general phenomenon. Within each group, the factor that contributed the most to the explanatory power of the regression model was retained. This process reduced the number of cost drivers to nine.

Finally, the analysts added to the regression model one at a time a few variables with the strongest theoretical basis for inclusion. Two were added to the final regression model.

In the end, the analysts concluded that seven factors should be included in the regression model explaining differences in the administrative cost per voucher. All factors had a strong theoretical reason for inclusion and the expected sign in all regressions. Four factors (wage index, program size, distance of voucher holders from PHA headquarters, percent of recipients with earned income) were highly statistically significant in all regressions and had estimated coefficients of a similar magnitude across many regressions. In the final regression that included only the seven factors and used a more refined measure of program size, the variable *small area rent ratio* is statistically significant, albeit with a coefficient very different from the earlier regressions. The other two variables included in the final regression (health insurance cost index and new admissions rate) had the expected sign but were less precisely estimated.

Other analysts using the same data to solve the same problem would undoubtedly have proceeded somewhat differently, and their results might have been at least somewhat different. For example, at the first step they might have chosen a preliminary set of factors based on how much each added to the fit of the regression model rather than the statistical certainty that the factor had some effect on administrative cost. However, the methods used in the Abt study were reasonable and systematic, and they led to a very credible formula for allocating the fees for administering the HCV program to the agencies that operate it. In my judgment, additional refinements are unlikely to produce enough of an improvement to justify their cost.

Finally, the methods and results are well documented and clearly explained.

Although the exposition in the report is generally quite good, it may lead the unwary to an unsupported conclusion, namely, that the proposed formula yields the administrative fee that a housing agency must receive in order to be high performing and efficient. This glosses over the reality that both performance and efficiency are matters of degree with no natural dividing line. The proposed formula is clearly much better than the current formula in accounting for differences in the cost of administering the program equally well in different settings. However, it does not indicate the appropriate *level* of administrative fees because that depends on the desired level of performance. The proposed formula reflects the difference in the average administrative cost of housing agencies in different settings among agencies that are classified as high performing according HUD's evaluation system and efficient according to the criterion adopted in the report. It is surely not true that all agencies that are classified as high performing have the same level of performance or that all agencies classified as efficient are equally efficient. The report contains information on the latter. Efficiency was measured as the number of vouchers per FTE staff. In the initial sample, the mean was 130 and the range was 25 to 300. Programs with more than 50 vouchers per FTE staff were described as efficient. Almost all programs are in this category; about 97 percent of the high performance agencies that agreed to participate were classified as efficient. The most efficient high performing agency was six times as efficient as the least efficient classified as efficient. Obviously, the study set a low bar for efficiency, indeed, so low as to raise questions about describing the agencies as efficient. If a higher bar for efficiency had been adopted, the formula would have generated lower administrative fees.

The preceding remarks are based on the reasonable presumption that greater administrative fees would lead to better performance, if not greater efficiency. The report contains no evidence on the effect of the administrative fee on performance or efficiency. Under the current system, the agencies that are not classified as high performing and efficient may receive as much in administrative fees as the agencies with the same cost drivers that are so classified.

Since the Peer Reviewers reviewed the Draft Final Report and the Abt Associates, Inc. team of researchers was working on completing the Final Report, Abt Associates, Inc. team was able to respond to the comments from the two industrial engineering experts, which appear below. Abt Associates Inc team responded to the comments from Dr. Ed Olsen in the Final Report text in track changes.

Edgar O. Olsen
Professor of Economics and Public Policy
University of Virginia

Response to Dr. Edgar Olsen's Comments, 17 June 2015:

I would like to highlight two of the changes that Abt made in response to your comments. First, Abt revised the definition of the small area rent ratio variable. Please let me know if this definition is clearer than the previous one. The definition is below and on pg. 159 of the Final Report.

"For PHAs in metropolitan areas, the small area rent ratio is calculated as the median gross rent for the zip codes where voucher holders live, weighted by the share of voucher holders in each zip code, divided by the median gross rent for the metropolitan area. For PHAs in non-metro counties, the small area rent ratio is calculated as the unadjusted two-bedroom FMR for the non-metropolitan counties where the PHA operates divided by the published FMR. Thus, the small area ratio is less than 1 if the county has a state minimum FMR, and equals 1 otherwise. The PHA's ratio is the average across counties served weighted by the share of the voucher holders in those counties."

Second, Abt also revised the discussion of data volatility, which is now on pgs. 185-187 of the final report. Hopefully that discussion is clearer as well.

There were some other changes, but they are harder to identify. It would be great to get your feedback on the two changes I highlighted above.

Final Peer Review Comments from Dr. Edgar Olsen, 18 June 2015:

The revisions in response to my comments on these two points are fine. It makes sense to use the average values of variables in the administrative fee formula over a few years to reduce year-to-year variability in administrative fees.