Evaluation Tradecraft

Evaluation Tradecraft presents short articles about the art of evaluation in housing and urban research. Through this department of Cityscape*, the Office of Policy Development and Research presents developments in the art of evaluation that might not be described in detail in published evaluations. Researchers often describe what they did and what their results were, but they might not give readers a step-by-step guide for implementing their methods. This department pulls back the curtain and shows readers exactly how program evaluation is done. If you have an idea for an article of about 3,000 words on a particular evaluation method or an interesting development in the art of evaluation, please send a one-paragraph abstract to marina.l.myhre@hud.gov.*

Improving Program Evaluation: Using Direct Time Measurement for Estimating Administrative Costs

Kevin Hathaway
RSG

Jennifer Turnham
Abt Associates Inc.

Abstract

Benefit-cost analysis is a common component of evaluation studies. Although techniques to establish the benefits continually improve, approaches to estimate costs based on how staff distribute their time remain antiquated, with imprecise timesheet instruments and precarious assumptions. For government agencies and researchers interested in accurately evaluating these costs, and with labor remaining the largest share of administrative spending for many programs, better techniques are needed for recording and measuring staff time. We present the direct time measurement approach from the Housing Choice Voucher Program Administrative Fee Study, describing the techniques and technologies used and discussing the logistical work required to ensure a successful time measurement effort.

Introduction

The Housing Choice Voucher Program Administrative Fee Study (HCV Study) was a multiyear economic evaluation that the U.S. Department of Housing and Urban Development (HUD), Office of Policy Development and Research, commissioned to ascertain how much it costs to administer a high-performing and efficient HCV program and to recommend a revised formula for allocating administrative fees to public housing agencies (PHAs) based on objective cost drivers. Because objective fee allocation would require a determination of administrative costs measured through staff time allocation while adjusting for agency and region-specific variables, a detailed analysis of HCV frontline activities was one of the study's main priorities. During the study's design period, HCV staff time at four PHAs on all program activities was evaluated using three test approaches: (1) traditional paper timesheets, (2) direct human observation, and (3) smartphone-based Random Moment Sampling (RMS). The research team determined that RMS was the most accurate for the level of detail required by the study and that it minimized staff burden and provided unparalleled research scalability compared with the other approaches. This article describes the RMS approach in detail and discusses the technology and staff resources needed to implement a successful multi-site time measurement effort using RMS.

Methodology

The HCV Study required a complex combination of research design, software development, and data collection logistics that evolved over several years. While describing the full extent of research design decisions that were needed or the technical details of building the underlying software frameworks are beyond the scope of this article, we briefly outline the key aspects that should be of highest use to researchers conducting similar evaluation studies.

Time Measurement Instrument Design and Approach

RMS, also known as activity sampling or work sampling, is a method for estimating the time spent on work activities based on randomly sampled data points collected during working hours over a period of time (Bolstein, 1986). Research goals, required time estimate precision, and the presumed a priori variance in work behavior dictate data-collection length and work-activity detail. Our approach issued random RMS notifications to participants via smartphone. Each notification contained an activity survey that required the participant to assign what they were predominantly doing for the previous 5 minutes to a work area and specific activity within that work area. For our purposes, we did not permit multitasking responses, so all answers required the participant to select one and only one activity. Participants submitted their responses to the notifications through a series of touchscreens on the smartphone.

During the early stages of the study, our team completed the following steps to design the time measurement instrument.

1. Reviewed differences in how identical work functions are identified by staff at the studied agencies and create a common language for all participants to reference.

2. Created a mutually exclusive and exhaustive list of HCV frontline activities and also categories for frontline work on other programs operated by the PHA and for overhead work, so all staff time could be assigned to an activity and to only one activity.
3. Provided accessible definitions in the smartphone app and supporting materials to help staff accurately assign their time.

The time measurement instrument included more than 50 frontline HCV activities, organized in a tiered structure so that participants started by identifying the program they were working on, then a main work area within the program, then an activity within that work area, and then (for some activities) a subactivity. Two reasons justify this level of detail. The first is that the study needed a certain level of detail to answer its research questions. For example, the study needed to quantify the time spent on eight special voucher programs (serving specific populations) and a related program (the Family Self-Sufficiency program) in addition to the regular HCV program. Within these programs, the study also needed to measure the relative time spent on six core program functions and specific activities within those functions. The other reason for the level of detail, however, is that we learned from testing the approach that the participating staff needed to see the specifics of their work reflected in the time measurement instrument to feel invested in the data-collection effort. Staff members who had highly specialized roles found that repeatedly entering their time under one or two categories was demoralizing. Some staff also felt that grouping their work under broader activity headings without permitting them to provide more detail implicitly devalued their work, which was not the study's intention.

Sampling

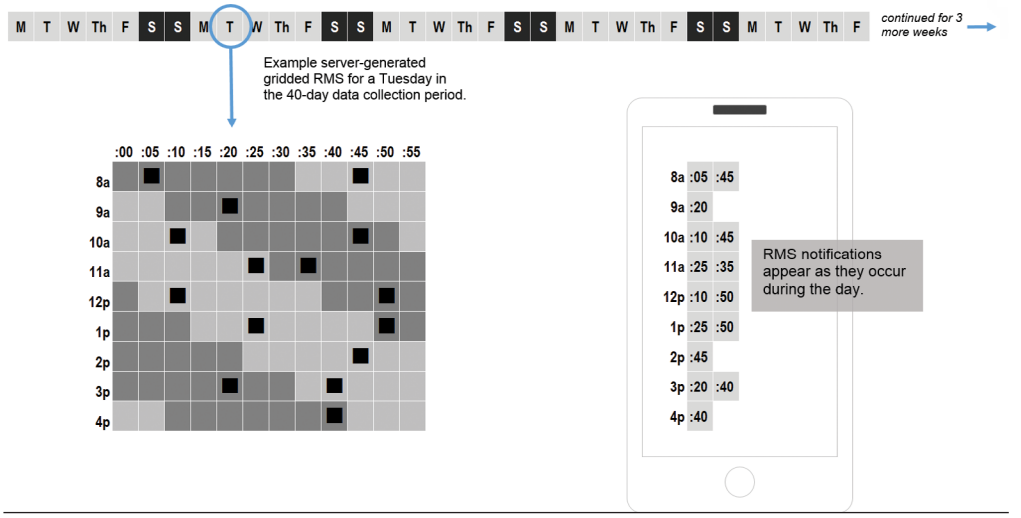
The RMS time measurement effort included 60 PHAs that all had a track record of several years of high performance as defined by HUD's performance measurement system for the HCV program, and the study team confirmed each PHA was high performing and efficient using onsite reviews. The 60 agencies participated in the RMS time measurement component in 12 waves covering nearly 18 months. Each agency participated for 8 continuous weeks. In selecting each PHA's 8-week timeframe, the study team balanced the need to conduct time measurement throughout the year with roughly equal cohorts of PHAs against individual PHA preferences for when time measurement would probably occur. We did not permit PHAs to select their 8-week period but tried to avoid times when key staff would be unable to participate because of long-term schedule leaves, such as maternity leaves or leaves of absence.

Every PHA staff member who worked on frontline HCV activities was included in the time measurement component at all but the largest PHA sites. At the largest PHAs, the study team selected a sample of staff to participate, with input from the PHA, based on the number of staff in each job category or work role. All staff who played a unique role in the program were selected for participation, but the team selected a random sample of staff for job categories—such as housing inspector—in which multiple staff were serving the same role.

At all sites, we conducted time measurement on participants for 40 working days during an 8-week period responding to 12 to 15 RMS notifications per day and tailored to each individual staff's work schedule (exhibit 1). Any nonscheduled night and weekend work was captured with

Exhibit 1

The 40-Day RMS Period With 1 Day of RMS Notifications for a User



RMS = Random Moment Sampling.

an additional sampling feature on the device. These irregular notifications persisted on the phone for only a few hours and then disappeared if left unanswered. In this way, the sample frame was partially dynamic, allowing for a person to work a long week and for the study team to capture that effort. This sampling plan was designed to detect small agency-level effects using power analysis with the arcsine transformation for differences in proportions (Cohen, 1988).

Technology

The team authored a native Android™ application, written in Java, that implemented a full RMS time measurement approach. Working with Verizon Wireless, more than 250 LG Vortex smartphones (devices) with network data plans were acquired for the data collection. Devices were provided to each participant in the first set of PHAs and then recycled to participants at subsequent PHAs to measure more than 900 participating PHA employees during the study. On average, each device was used by four different participants over the course of the data collection, for 32 weeks of daily use.

An intelligent web service (written in the Ruby on Rails framework) and secure database (MySQL) ran on a Linux web server. The web service was designed so the team could add participating PHAs and define their data-collection dates, add the roster of staff members and their unique work schedules, and add specific settings that affected how the RMS was conducted for each participant. The web service then autogenerated a full sampling scheme for a participant with pregenerated times, which would be synced to the participant's issued device for their 8-week period. The user was blind to all pending RMS notifications, which showed as an unanswered work survey displayed on a simple daily calendar in the app. Notifications used the device's native ringers or

vibrate setting, which could be easily changed by the participant. Answers were streamed instantly to the web server, assuming a data connection existed. Automatic database backups and the ability to launch a replicate web server and database were implemented in the event of a server problem.

Our team designed the Android app and web service to work together but to continue functioning correctly if data connections were lost, particularly for those PHAs located in rural communities. For example, if 3G service was lost as a housing inspector drove to a rural location, the app would continue to issue notifications at the predetermined times. Without the data connection, the answers were cached locally on the device until the data connection was restored. The web server, at the same time, reported the device had gone "dark" because the app and server were designed to communicate every 5 minutes. No action was taken by the study team unless a device remained dark for more than 1 day.

RMS Training

Although smartphone data collection was relatively easy for PHA staff, it still required training and ongoing support. Each of the 60 PHAs participating was assigned to two research team members serving as site liaisons. We also asked each PHA to designate a staff member to serve as the PHA liaison to troubleshoot any problems and encourage staff to answer notifications in a timely manner.

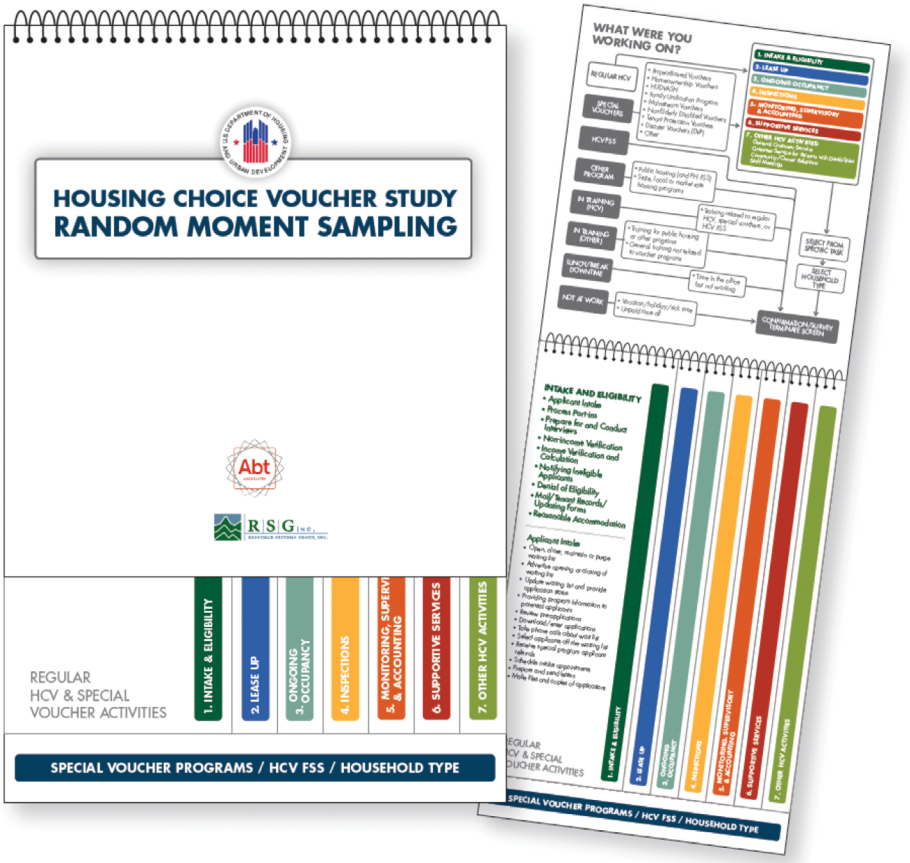
Research site liaisons conducted an in-person, 2-day training event at each PHA. On the first day of the training, the research team met with PHA staff to introduce study goals and to demonstrate how to use the smartphones and how to classify their work using the HCV activity and subactivity categories. At smaller agencies, all staff were trained together. At larger agencies, the training was grouped by staff function (for example, inspectors were often trained together). The staff began responding to notifications immediately after the training, although we did not use the first day's data in the analysis.

On the second day of training, the site liaisons met individually with each participating staff member (or in groups of two to three at the larger sites) to address questions and determine what specific categories and subcategories staff would probably be using for their own work functions. These sessions were highly productive because they came after staff had had an opportunity to practice with the device the previous day. These individual meetings also gave the site liaisons an opportunity to take detailed notes on individual staff members' HCV activities so they could monitor the accuracy of their responses during the 2-month period.

To reinforce the training, the team provided all staff with a 6- by 8-inch spiral-bound training booklet, with easy-to-read tabs, describing the response categories that staff would see on their smartphones. Under each response category, the training booklet included a listing of all the possible activities and tasks that could fall under that category, so that staff could readily find the right category for the particular task they were working on (exhibit 2). The team updated the training booklet at several points during the study with new tasks under the response categories based on feedback from participating staff. As mentioned previously, it was very important for accuracy and staff member motivation that staff be able to see the specific work they do reflected in the training materials. The Android app contained similar helper buttons but had less detail than the booklet.

Exhibit 2

Random Moment Sampling Training Booklet



Monitoring Data Collection

The team established several methods for monitoring RMS participation, including ongoing communication with the PHA liaison, a messaging system through the smartphone for participants' questions, and monitoring of the PHA staff responses to RMS notifications through a shared website dashboard. One benefit of using the smartphone technology was that all notifications were continuously uploaded to a central server where the study team could view them virtually in real time. For each PHA, the website included information on staff participating in RMS, their work schedules, how many RMS notifications were outstanding, their actual answers, how long after each notification the user answered the RMS survey, all text messages sent to and from each staff member, and the current battery power of each RMS device. The study team dedicated one staff member to monitor the website dashboard and incoming messages during working hours. The study team developed a series of decision rules regarding when to contact PHA staff if it looked as if they had stopped responding to their notifications. The team also made extensive use of the app's custom messaging function, which worked like a text message, to send reminders to participating staff.

Estimating Time and Costs

This systematic surveying of activities for the sampled agencies returns several needed pieces of information. The first is a count of notifications assigned to mutually exclusive HCV functions. The second is the total estimated time staff worked during regularly scheduled hours and any time they worked outside those regularly scheduled hours. Because RMS sampling was designed to grid the notifications within the 36-minute blocks, or approximately 1.66 notifications per hour, the resulting activity counts were converted into total minutes of activity in the data-collection period. This sampling was completed using time expansion with appropriate sampling weights, because the RMS notifications were drawn with known probability. Simultaneous confidence intervals were computed using several methods, depending on the aggregation level, including Wilson Scores and intervals based on bootstrapping, but a number of alternative methods can perform well with multinomial data (Efron, 1987; Glaz, 1999). Calculating activity-specific costs for the full-time HCV employee is a simple extension of the computations using the resulting activity-level time distributions. We calculated overall HCV program costs and costs for each activity within the HCV program by multiplying the time spent that each individual staff member spent on the program and component activities by that staff member's salary and benefits. This calculation produced a direct labor cost for each PHA, to which we added nonlabor costs and a share of overhead costs. (We collected information on nonlabor and overhead costs directly from each PHA through a combination of interviews and reviews of financial documents.)

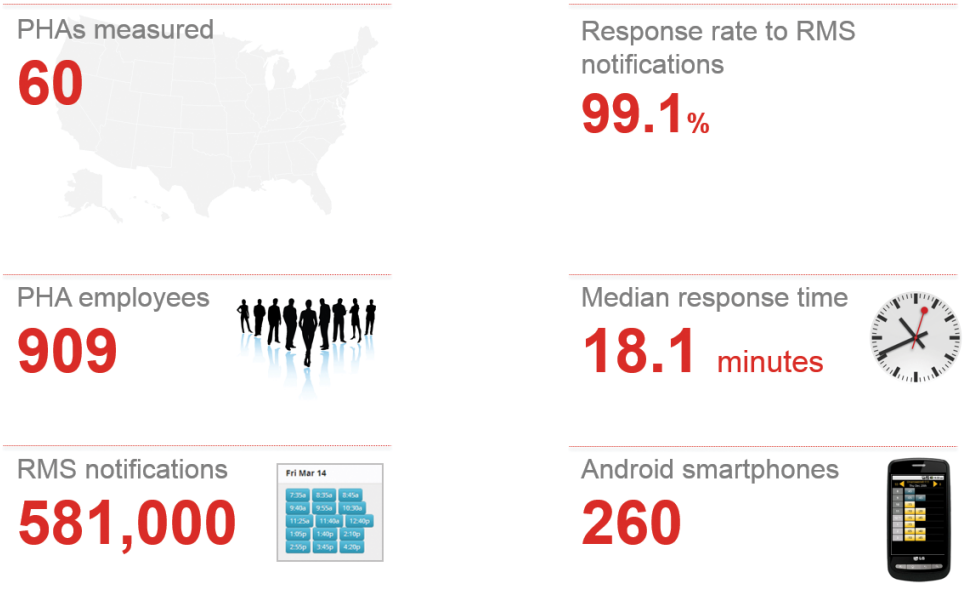
Discussion

Because data quality was a high-priority goal, the data-collection instrument, training materials, and monitoring work reflected that goal. For example, to simplify the monitoring tasks, the web service generated summary reports and automatically e-mailed the study team every 2 hours. The report included a list of PHA employees with outstanding notifications, a list of users who were waiting too long to answer the RMS notification (that is, response time), notification of an employee who was assigned the same activity more than 10 times in a row, notification of the server losing data communication with any smartphones, and notification of battery levels if they dropped below 5-percent power for any device. These reports enabled the liaisons to investigate further, using a series of interactive data visualizations (programmed using the D3 library for web development) on the shared website, using customized tools, or communicating with the participant when necessary. Team members communicated to any or all participants directly using an online text messaging service. Text messaging with participants proved invaluable to keeping participants answering all RMS surveys, doing so quickly, and assigning their work accurately. In the end, approximately 581,000 RMS notifications were answered, resulting in a 99-percent response rate across all staff (exhibit 3).

Response time in smartphone-based RMS is the amount of time after an RMS notification is issued that the respondent actually answers the notification. Minimizing response time addresses some of the recall bias issues plaguing so much self-reported data. Although keeping response times low was a primary goal, we expected response times to vary by work activity. For safety reasons, we instructed inspectors never to answer an RMS notification while driving. Similar guidance

Exhibit 3

Overview of RMS Data Collection



PHA = public housing agency. RMS = Random Moment Sampling.

was provided for participants while in staff meetings and meetings with voucher holders, because answering a smartphone survey could be considered rude in both contexts. As a result of this guidance, RMS answer categories for "driving to/from inspections" and "staff meetings," showed longer average response times than other core activities. Across all HCV activities, the median response time was 18 minutes. This median value indicates that of the 581,000 RMS notifications, approximately 290,000 were answered faster than 18 minutes from when they were issued.

To ensure the RMS responses were accurate and correctly assigned, the study team continually reviewed staff responses and compared those answers with the staff member's assigned work areas, as provided by the PHA liaison and by the participant during the second day of training. Any inconsistencies were confirmed with the participant via messaging. For example, if a housing specialist who primarily worked on annual recertifications responded to a notification claiming she was conducting an inspection, the team confirmed with her that the response was accurate. If the study team detected any unusual patterns in responses that could indicate trouble in understanding the reporting categories, we contacted the PHA liaison or staff directly to retrain on the HCV activity definitions. In a small number of instances, staff had systematically assigned work to the wrong category. Several features were available in the web system to enable the research team to reassign those specific answers to the correct category. Participants could also edit their own RMS responses for up to 24 hours via the touchscreen. In all cases, the original answers and any edits were maintained in the database for the study team's records.

In addition to providing the remote monitoring, we provided the PHA liaison with a report on the RMS responses in aggregate for all participating RMS staff and their overall median response

time after 1 month of data collection had passed. This midpoint report was another opportunity to detect any inconsistencies in reporting and it served as a motivator to staff to continue their timely responses to notifications.

The primary advantage of using RMS on a smartphone is data quality. Before mobile computing, RMS methods were typically conducted by supervisors recording their employees' work activities on a clipboard at predetermined and randomly drawn times. Eventually the methodology was migrated to mediocre software on the desktop computer, requiring a user to be at a computer to respond. With the advent of the smartphone, many of the historical challenges are gone and the burden drastically reduced for both the evaluators and the participants. Further, the ability to monitor and communicate easily in real time has further elevated its application and usefulness in modern evaluation studies.

The study team also gained operational efficiency by using tools built to support the research goals. Because field data collection always carries some level of uncertainty, having technologies facilitate flexibility can be crucial. When the team arrived at a PHA to start data collection, it was common to discover additional staff needed to be included in the RMS work or that participating staff had work schedules that differed from what the PHA had previously indicated. In only a few minutes, all changes could be made using the web service—the sampling scheme instantly regenerated and the updates were pushed to the appropriate smartphone. If a device was lost or broken at any time, spare devices left at the PHA were swapped in with only a few touches. The lost or stolen devices were remotely locked, so they became useless. These and many other operational research functions were made easier by using the described technologies.

Caveats and Conclusions

Using a technology-based RMS approach generated a vast amount of high-quality data. Estimating the labor component of administrative costs from RMS, however, relies on several assumptions.

- Participating staff must accurately answer the RMS surveys and correctly assign their work at the selected time to the proper RMS activity category. To achieve this accuracy, researchers must build (and test) strong data-collection and training tools and have a means for monitoring responses and communicate with participants in real time.
- In the case of normalizing time across different participating agencies, the denominator (normalization variable) must be known without error. For example, if you wish to define time per housing inspection, then a separate data-collection effort to ascertain the number of inspections that took place during the time measurement period is required. (This calculation was done for the study by requesting counts of important transactions from each PHA for the study period and the preceding 12 months.)
- The resulting statistics describe activity time during the RMS period only and may or may not represent longer term patterns of work. Alternative RMS designs are available to address this issue, such as extending the data-collection period or repeating the data collection at several points during the year. These alternatives carry varying levels of logistical burden to the evaluation and to participating staff.

Evaluation studies can be expensive, and relying on technology-based methods for collecting data carries measured risk. Network reliability, undetected software bugs that emerge in the middle of data collection, catastrophic server failures, or lost data can be expensive to remedy. Because of these risks, systems must be designed thoroughly, with as much protection as can reasonably be included. As the use of technology continues to invade all aspects of modern research and evaluation, so too will the expectations for how well these systems should work. The public is increasingly exposed to great technologies offered by today's largest companies; researchers offering their own solutions that are not as easy to use, are less reliable, and are not as robust may experience challenges.

For those researchers interested in similar approaches, we recommend thinking about the scale of the work and if technologies will be sufficiently leveraged. When data-collection efforts become large enough, the benefits of technology become more obvious and the economics become more attractive. Off-the-shelf options that are also emerging may satisfy a large number of simpler evaluation studies in the near future.

Acknowledgments

The authors thank their key project team members at RSG and Abt Associates and also the participating public housing agency employees for their dedication to the Housing Choice Voucher Program Administrative Fee Study. They also thank Marina Myhre at the U.S. Department of Housing and Urban Development, Office of Policy Development and Research, for her helpful feedback on drafts of this article.

Authors

Kevin Hathaway is a senior researcher and the Vice President of Quantitative Methods and Advanced Technologies at RSG.

Jennifer Turnham is a senior associate in the Social and Economic Policy Division of Abt Associates Inc.

References

Bolstein, Richard. 1986. "Random Moment Sampling To Estimate Allocation of Work Effort." In *Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association: 671–675.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlabaum Associates.

Efron, Bradley. 1987. "Better Bootstrap Confidence Intervals (With Discussion)," *Journal of the American Statistical Association* 82 (397): 171–200.

Glaz, Joseph, and Cristina P. Sison. 1999. "Simultaneous Confidence Intervals for Multinomial Proportions," *Journal of Statistical Planning and Inference* 82 (1999): 251–262.