AN EXPERIMENT WITH PAYMENTS TO SURVEY RESPONDENTS

Stuart H. Kerachsky, Charles D. Mallar, and Jewel Moran, Mathematica Policy Research[1]

This paper summarizes early findings from an experiment with payments to survey respondents for their time and effort in being interviewed. In addition to considerations of equity, such payments have been advocated for surveys to reduce search effort, increase response rates, and improve data quality through hypothesized effects on interviewer efficiency and respondent cooperation. Furthermore, the effects of respondent payments are often thought to be most important for longitudinal surveys with repeated waves of interviews and for populations that are difficult to locate and interview.

The payments experiment discussed here was conducted in conjunction with a national evaluation of the economic impact of the Job Corps program, which provided a sample where respondent payments were expected to be effective. The sample includes a longitudinal panel of nearly 4,400 observations of economically disadvantaged youths, approximately 70 percent of whom were offered $5 payments.

Although the literature on the use of respondent payments in surveys does not clearly establish their value, there is no convincing empirical evidence with which to reject the efficacy of payments. Offering payments is supported by their hypothesized effects in three general areas of survey operations. The first hypothesized effect concerns interviewer efficiency in the search effort. For example, interviewers claim that they can operate more effectively if they are able to offer payments, because they can more easily engender credibility with community members, whose aid is sought in the search process as well as with potential respondents. In addition, there is expected to be an impact from the feeling among interviewers that payments offer respondents a fair deal--compensation in exchange for time and effort, thus making the interview seem less of an imposition on respondents.

The other two hypothesized effects relate more directly to the reactions of potential respondents; thus, they have more serious implications for the quality of the data. The second effect concerns cooperation in an initial survey: without payments, particular segments of the target population may be underrepresented in the sample because they are more likely to not want to be found and will be less likely to agree to cooperate if they are found. Furthermore, those who are interviewed may not give the attention required when responding to detailed questionnaires. The third effect concerns cooperation in a follow-up survey: both the tracking procedures for subsequent follow-up interviews and the cooperation of sample members at the time of the interviews may be facilitated if respondents can anticipate compensation for their assistance. Thus, with repeated interviews, payments may mitigate both sample attrition and item nonresponse. This hypothesized effect becomes even more important when long-term follow-ups are anticipated.

While beneficial effects from respondent payments were certainly plausible, not much was previously known about the magnitudes of these effects (the empirical research had been limited to small scale experiments that yielded little precision for estimates). Because of the uncertainties, payments were initially advocated as a prudent approach for the surveys undertaken in conjunction with the evaluation of Job Corps. However, it was subsequently decided that the expected benefits from more information on the effectiveness of respondent payments outweighed the potential risks to the project in terms of survey costs and data quality, and an experiment was undertaken to estimate the effects of offering payments to survey respondents. The design of the experiment and our findings from the first two waves of interviews are summarized in the sections that follow.

A. DESIGN OF THE EXPERIMENT

The Job Corps evaluation will include at least three waves of interviews with a sample of Corpsmembers and a comparison sample. The payments experiment was not a factor in the overall sample design for the Job Corps evaluation, both because it was a secondary objective and because federal clearances for the payments experiment were received late. Therefore, our task was to fit the payments experiment into a sampling strategy that was otherwise fully implemented.

Baseline Design. The Job Corps evaluation consisted of two areal probability samples--one for Corpsmembers and another for the comparison group.[2] The Job Corps sample covered approximately one-third of the three-digit zip-code areas in the continental United States that were randomly chosen to provide a probability sample of approximately one-third of the Corpsmembers at centers in May 1977. The comparison sample was drawn in fifteen geographic areas scattered throughout the nation--ten urban and five rural --and the comparison sites were selected to be similar to the areas from which Corpsmembers came, except that the sites were in areas where Job Corps did not recruit heavily and were distant from Job Corps centers. Within comparison sites, youths were randomly selected from recent lists of school dropouts and employment service applicants that were stratified to yield youths similar to Job Corps participants, on average, in terms of age, education level, length of time out of school, race-ethnicity, poverty status, and other socioeconomic characacteristics.

Only the comparison group was included in the payments experiment for the baseline survey, because Corpsmembers were still at centers and being paid by Job Corps. Hence, baseline payments to Corpsmembers would probably have been unnecessary and disruptive for center operators. Because of the lateness of the decision to undertake the payments experiment, even the comparison group design was somewhat limited at baseline. For example, no advance letters could be sent to sample members to notify them of the payments.

The experimental design for respondent payments was developed so that comparison sample members were randomly assigned to payment or nonpayment status by geographic areas. (Thus, within each area, either all of the respondents received payments for their interviews or none of them did.) This areal approach was implemented for two reasons. First, within-site assignments would likely result in sample contamination from information being spread to some of the youths in the nonpayment group that others were receiving payments, which would, in turn, likely bias the nonpayment group toward refusals. Second, in order to be able to engender community support when attempting to locate respondents, an area saturation was desirable. Because previous experience had indicated that urban and rural areas pose very different interviewing problems (particularly with regard to locating sample members), we decided further that these two types of sites needed to be kept separate for the baseline phase of the payments experiment.

The next major task was to determine the proportion of sites to be assigned to payment status. The experiment with payments would have best been served by assigning 50 percent to payment and 50 percent to nonpayment status. On the other hand, the risk-minimizing option for the primary objectives of the Job Corps evaluation would have been to assign all sites to payment status. Our resolution of these conflicting objectives was to assign 70 percent to payment and 30 percent to nonpayment status, which gave adequate power to detect large payment effects but did not seriously jeopardize the main evaluation. For the urban sites, this obviously suggested seven payment and three nonpayment sites. For the rural areas, the split had to be either three and two, or four and one. Because we expected that there would be substantial variance in responses among rural sites, neither of these rural distributions would have yielded very precise empirical evidence. As a result, we adopted the position of providing payments to everyone in the rural sites. Thus, the baseline experiment with payments was applied only in the urban sites.

The three nonpayment urban sites were selected in a random draw, subject to two constraints. First, one site had to be included in the payment sample because of previous agreements that had been made with local school officials prior to the selection process. This agreement also required us to mail letters to potential sample members to give them advance notice of the study. The additional contact with the sample at this site may have increased our ability to locate sample members, and this must be considered in the evaluation of the payments experiment at baseline. The second constraint was geographic in nature: a simple regional stratification was implemented to improve the efficiency of estimates. Altogether, approximately 300 of the 1,000 comparison-group youths in urban areas were not to be paid for their time and effort in taking the interview (the sample sizes for potential respondents were nearly 40 percent larger).

First Follow-Up Design. For the first follow-up (nine-month) and subsequent surveys, the payment status assigned to the comparison sites was to be the same as that assigned in the baseline survey. However, a payment status still needed to be randomly assigned to the youths in the Job Corps sample. Therefore, a random assignment procedure was implemented to select the sites where Corpsmember respondents would be paid for their time and effort. The three-digit zip codes of baseline residences were used as the probability sampling units. Cities and their suburbs were clustered together, where appropriate, both to minimize the chances that respondents not receiving payments would find out that others were being paid and to maximize the effectiveness of payments on community support. For these same reasons, we again decided to use 100 percent sampling rates within sites for payment status (i.e., either everyone in a site received payments or nobody did).

The payment status for the members of the comparison group was assigned at baseline, as discussed above, and the remaining Job Corps sites were chosen so that payments would be made in approximately 70 percent of the Job Corps sites. For the follow-up sample of Job Corps participants, 951 of the 2,892 Corpsmembers on the follow-up sample list would not be offered payments. For the payment experiment, the country was effectively divided into 186 areas (primary sampling units identified by zip code); 3,126 youths (ultimate sampling units) in 129 areas would receive a $5 payment, and another 1,262 youths in 57 areas would not receive a payment for taking the interview. Thus, the follow-up sample sizes for the payments experiment are much larger than for the baseline survey, and should result in more precise estimates.

The rural-urban distinction was not maintained for the follow-up sample of Corpsmembers because (1) the large number of geographic areas ensured adequate representation of payment and nonpayment assignments in both rural and urban settings, and (2) many of the areas contained both urban and rural subdivisions. With the different distributions of urban-rural mix and payment statuses between the Job Corps and comparison groups, care had to be exercised in distinguishing payment effects on response rates from Job Corps effects.

While the payments experiment began too late to fully exploit payments in the sample search effort for the baseline survey, they were more fully utilized in the tracking and field searches for the first follow-up interviews. The advance mailing to potential respondents notified sample members in all payment sites that they would be paid. (Sample members in the nonpayment sites received similar letters, but with no mention of payments). In addition, members of the comparison sample in the baseline payment sites were told at baseline that they would again receive $5 payments when they were reinterviewed after nine months. When obtaining assistance from local persons in attempting to locate hard-to-find sample members, interviewers explained to them that their cooperation would result in a monetary benefit to the sample members. Finally, interviewers attempted to use the payments to gain the cooperation of the respondents once personal contact had been made.

B. FINDINGS FROM THE BASELINE SURVEY

We knew from the outset of this experimental study that the baseline data would not provide very powerful tests of the effectiveness of respondent payments. More important than the problems caused by the late start-up discussed above (e.g., from the lack of notification of payments through advance letters), the sample size was small and highly clustered. In addition, some of the hypothesized effects of respondent payments relate to maintaining the sample over repeated waves of interviews. Given these limitations, the baseline data were expected to provide only an indication of potential effects that would help support the more powerful statistical tests provided by the follow-up surveys when the Job Corps sample would be included in the payments experiments.

Two comparison-group sites have to be given special consideration when analyzing the baseline data. As explained above, one site had to be included in the payment sample because of agreements we made with local school officials prior to the selection of payment sites. Another site presented a different type of problem for the experiment at baseline. Here we encountered an unusually large number of bad addresses on the sample lists, which adversely affected the search performance. The unusual number of bad addresses was complicated by a short list of school dropouts which resulted in an oversampling from recent applicants to employment service offices. Addresses obtained from employment service offices for disadvantaged youths were often outdated or nonexistent in all of the comparison sites.

The effects of payments to respondents were evaluated through differences in both the success of the search effort and the quality of interview data. The former includes the ability to find sample members, interview completion rates per potential sample, the refusal rates per contact, and the person hours spent searching per completed interview. The latter includes the numbers of "Don't Know" responses, refused answers, blank answers, and total item nonresponses per completed interview. The latter also includes the length of completed interviews in minutes--a variable that may reflect either the amount of thought that respondents gave to their answers or their degree of attentiveness (i.e., a priori arguments could be made for either a positive or negative impact of payments on the length of interviews).

The findings are based on comparisons of sample means among payment and nonpayment sites. The results are summarized in Table 1. The first column contains the variable means for the three urban nonpayment sites. The second column provides a direct comparison with the means for the seven urban payment sites. The last column contains our best estimates of the means for payment sites. This differs from the second column in two major respects. First, the two problem urban sites that are discussed above were not included in the analysis of the search- and completion-rate variables. Second, the best estimate for the quality of interview measures includes <u>all</u> payment sites (all seven urban and all five rural sites). The separation of the urban from the rural sites was based largely on search and completion problems, and not on the quality of interviews. Once an interview is underway, there is no obvious reason to distinguish quality by population density. Similar arguments support including the two problem urban sites when estimating payment effects on the quality of interviews.

The effects on interview completion rates and person hours spent searching per interview were inconclusive, which is not surprising given the sample size and other limitations of the baseline payments experiment. For example, the effects of payments may have been artificially low in the baseline study because the clearance for payments was received too late to send out letters notifying sample members about the payments. As hypothesized, the completion rates were higher in the payment sites (significantly so for the best estimate). Counterintuitively, however, the person hours of search per completion were somewhat higher in the payment sites, which may partly explain the higher completion rates. Also, the fraction of refused interviews was slightly higher in the payment sites (insignificantly so), which is counterintuitive.

In contrast to the first set of variables, there is stronger evidence that payments to respondents did indeed result in higher quality answers for those interviewed. Among those who were given payments, the interview was slightly longer (significantly so), there were fewer refused answers per interview, and there were fewer "Don't Know" answers per interview (significantly so). However, the total item nonresponse differences were somewhat offset by inexplicable but statistically significant differences in the opposite direction for blank answers per interview, which are a combination of interviwer error (e.g., forgetting to fill in answers and illegible writing) and uninterpretable answers (e.g., to questions on occupation). It is difficult to see how payments would affect the number of blank answers, especially because it can be shown that they resulted primarily from interviewers' forgetting to indicate that they had written asnswers to confidential questions on a separate form.

The total number of unanswered questions (overall item nonresponse) was lower on average for respondents who were paid--approximately one fewer unanswered question per interview (but not statistically significant). This finding, combined with the results for the length of the interview, provides some support for the hypothesis that respondents who are given payments provide more thoughtful and accurate answers.

The payment experiment during the baseline phase provided only limited evidence for benefits from paying respondents for their time and effort in taking interviews. The only trend that emerged from the baseline interviews of the comparison sample was that sample members who received payments seemed to provide more complete and thoughtful information. However, there was some confliciting evidence, and better estimates of these effects can be obtained from the data for the first follow-up survey.

C. FINDINGS FROM THE FIRST FOLLOW-UP SURVEY

In contrast to the baseline survey, the first follow-up survey was conducted with more effective use of payments for sample members in the payment sites. In practice, this meant that the use of payments in the field search and interview process included the promise of payments to the appropriate sample in advance letters that were sent to sample members to solicit up-dated addresses. Thus, an analysis of the first follow-up survey results should more accurately show the magnitude of benefits from offereing $5 payments to disadvantaged youths. In addition, the effective sample sizes were much larger at follow-up than at baseline--more observations, much less geographic clustering, and a little more even spread across payment and nonpayment statuses--which will yeild more powerful statistical tests.

The analysis, itself, differs in three respects from that with the baseline data. First, while the same measures of data quality are analyzed (the various measures of item nonresponse and length of interview), the measures of search efficacy and interview completion are changed to be more appropriate to the special features of the follow-up. The measures that were examined are the fractions of the sample that (1) returned postcards in response to the advance letters for address updates, (2) were located (whether or not they were interviewed), and (3) provided at least some analytical information, and (4) refused to be interviewed if contacted.

The second difference in the analysis of the baseline and the follow-up surveys stems from the change in the sample. As was noted earlier, the analysis of the baseline survey was confined to the comparison group; moreover, even within the comparison group, the rural sites and two urban sites had to be excluded from some of the analysis. For the analysis of the follow-up data, the full sample of observations can be included.

The final difference is in the mode of analysis. Like the baseline analysis, the follow-up data could be analyzed through a basic comparison of sample means. However, unlike the baseline analysis, the availability of baseline data for everyone in the follow-up sample permits regression analysis to control statistically for individual differences and to detect any payment effects that are specific to certain subgroups. However, with a classic experimental design, the variable representing the experimental treatment will tend to be orthogonal to the other variables (unless some unusual distribution is obtained despite the randomization); hence, sample mean comparisons will tend to be similar to regression estimates, unless the other variables included explain a large portion of the variance in the dependent variables and improve the t-statistics. The other variables that we included in regressions controlled for age, education, race-ethnicity, mobility, marital and family status, arrest history, health, sex, baseline earnings, and Job Corps-comparison group status. These variables generally did not explain very much of the variance in the dependent variables of interest (i.e., often less than 1 percent), even though some of them were statistically significant.[3] Thus, the findings from comparisons of sample means are emphasized in this paper.

The main findings from the analysis based on a comparison of sample means are summarized in the last two columns of Table 1. The first of these two columns shows the sample means for the nonpayment group for each of the several outcome variables; the last column shows the sample means for the payment group.

The results for search efficacy and interview completion generally show very small impacts from payments to respondents. The largest effect found for search efficacy is the return of postcards from the advance mailing to verify or update addresses. The sample mean differences between the payment and nonpayment groups shows an increase of just over one returned postcard for every 100 sample members in the payment group, from a base of approximately 27 returned postcards per 100 sample members. This effect is small and statistically insignificant but becomes larger and more significant in the regression estimates. By simply controlling for Job Corps versus comparison status, the effect increases to nearly three additional returned postcards for every 100 sample members in the payment group, with a statistical confidence of over 90 percent for a one-tail test (over 85 percent for a two-tail test). These results show some potential benefits to payments both in ease or cost of locating potential survey respondents and completion rates for mail surveys.

The next two variables (percent of sample located and percent with at least some final status information) measure the ultimate success of the search effort. Eight-seven percent of the sample was located, and we obtained some information on 85 percent of the sample. (The 2 percentage-point difference was due to the small number of refusals.) The payments had no effect on the ultimate success of the search effort, with the small, insignificant differences indicating slightly greater success in the nonpayment sites. For the first follow-up survey, it appears that we could have been equally successful in locating our sample of disadvantaged youths without the aid of respondent payments.

We had very low refusal rates for the first follow-up survey (only approximately 2 out of every 100 sample members contacted refused to be interviewed). Payments to respondents appear to have reduced this low refusal rate only slightly (just over one fewer refusal for every 200 sample members contacted). This effect was not significant for the sample means comparison, but became more significant in the regression estimates. While the effect on refusal rates did not seem to matter much for our sample (partly because of the extremely high level of cooperation), the marginal significance indicates that payments could be an important factor in surveys with higher overall refusal rates.

The effects of payments on data quality are shown by the last six variables in Table 1. These results reinforce the evidence found in the analysis of the baseline data that payments to respondents result in higher quality answers among those interviewed. The most notable effect is for the number of "Don't Know" responses to questions, which has a mean of nearly one per interview for the nonpayment group, but only slightly over one per every two interviews for the payment group (a statistically significant difference). On average, the payment group had nearly one fewer "Don't Know" response for every two interviews. This result may suggest that respondents who receive payments give more thought to difficult questions. Furthermore, the questions with the most "Don't Know" responses tended to be the most important questions for the evaluation--that is, questions on employment and other aspects of economic status.

Compared with nonpayment-group interviews, payment-group interviews also had fewer refused and fewer blank answers on average. However, neither of these differences is large or statistically significant. The payment effect on the total number of item nonresponses per completed interview--the cumulative effect on the "Don't Know," refused, and blank answers--is relatively large (nearly one fewer item nonresponse per interview), and the effect is statistically significant.

Contrary to the baseline findings, the first follow-up interviews for the payment group took less time to complete, on average, than interviews with the nonpayment group. The sample mean difference is very small (less than one minute, or 3 percent, of the interview time for the nonpayment group), but it is marginally significant in the statistical sense (greater than 90 percent confidence). However, with the regression estimates, the difference narrows and is no longer statistically significant. Therefore, the findings on length of interview are not clear-cut for the follow-up survey, but they do not suggest that respondents were taking more time to consider questions, as had been suggested by the baseline data.

The regression approach was taken one step further in order to investigate whether the payments to respondents were effective for any particular types of sample members. The experimental-payments variable was interacted with all of the other variables included (representing several socioeconomic characteristics of the youths). Statistically significant effects were rarely found with any of the payment interactions for any of the dependent variables. The few significant effects that were found formed no pattern and appeared to be random. Thus, we have been unable to identify subgroups of disadvantaged youths for whom the payments are significantly more or less effective than the average.

While the regression approach adds little to the analysis of the effects of payments to respondents, it can serve to focus our attention on other sample characteristics that affect the outcomes of interest. Further investigation is planned of the patterns for variables other than payments, which should facilitate the design and implementation of better survey methodologies and show more clearly any limitations in the current data.

D. CONCLUSIONS

This paper summarizes the evidence to date from an experiment in progress designed to test the effects of paying survey respondents for their time and effort in taking interviews. The main benefit from respondent payments is a reduction in item nonresponse for respondents who take the interview. For the payment-treatment group, there are fewer "Don't Know" answers to questions than with the nonpayment-control group.

After one nine-month follow-up survey, respondent payments have not enhanced our overall ability to locate and interview potential respondents. However, there is some evidence that respondent payments reduce the search effort by inducing sample members to be more conscientious in replying to advance letters that solicit a verification or update of their addresses. The follow-up survey also shows a marginal reduction in the number of refusuals to take the interview with the payments. However, the base for refusals is very small in our sample, and a slight change in a small base is not of much overall consequence.

Care must be exercised in extrapolating the findings presented here beyond the context in which they are estimated. Our experiment is directly applicable only to surveys with disadvantaged youths which use intensive search procedures, and thus far we have primarily been examining in-person interviews. We cannot predict very well what the effects of respondent payments would be for populations other than disadvantaged youths. The evidence that does exist suggests that the beneficial impacts of respondent payments could be much larger for surveys with high nonresponse rates caused by potential sample members refusing to be interviewed. The findings also suggest that respondent payments could lessen the nonresponse problems with mail surveys.

Some interesting and important findings are presented in this paper. However, most of the findings are, by necessity, tentative at the current point in the experiment. The effectiveness of respondent payments in reducing sample attrition over time with repeated interviews can be tested much more powerfully with the next (i.e., the second) follow-up survey. The second follow-up survey will be nearly two years from the time of the baseline survey, and it will be the third interview for most respondents. The sample size will also be larger because some Corpsmembers who were still in the program at the cut-off date for the first follow-up survey will be out long enough to be interviewed productively for the second follow-up survey (it will be only the second interview for this portion of the sample). Finally, the second follow-up survey is going to rely much more heavily on telephone interviews, so any specific effects for telephone surveys can be tested with greater statistical power.

FOOTNOTES:

[1] We gratefully acknowledge the research assistance of Patricia Lapczynski and the editorial assistance of Thomas Good. The research reported in this paper was funded under a contract with the Office of Program Evaluation of the Employment and Training Administration, U.S. Department of Labor. Researchers undertaking such projects under Government sponsorship are encouraged to state their findings and express their judgments freely. Therefore, points of view or opinions stated do not necessarily represent the official position of the Department of Labor.

[2] For more details on the Job Corps and comparison-group sample designs for the evaluation, see the Interim Report (September 1977), the main volume of the First Follow-Up Report (December 1978), and Technical Reports A, B, and C of the "Evaluation of the Economic Impact of the Job Corps Program," Mathematica Policy Research, Princeton, New Jersey.

[3] For the binary dependent variables appropriate modifications of the regression approach were used (i.e., probit techniques).

TABLE 1

COMPARISONS OF SAMPLE MEANS AMONG PAYMENT AND NONPAYMENT SITES[a]

	Baseline Survey			Follow-Up Survey	
	Nonpayment Sites[b]	All Urban Payment Sites[c]	Best Estimate for Payment Sites[d]	Nonpayment Sites[e]	Payment Sites[f]
Fraction of sample that returned advance postcards[g]	n.a.	n.a.	n.a.	.266	.277
Fraction of sample located	.714	.731	.774**	.875	.867
Completed interviews per potential sample	.696	.707	.751**	n.a.	n.a.
Fraction of sample with same analytical data	n.a.	n.a.	n.a.	.856	.853
Refused interviews per interviewer contact	.025	.033	.029	.023	.017
Person hours searching per completed interview	3.999	3.983	4.330	n.a.	n.a.
"Don't Know" answers per completed interview	3.873	3.263***	3.119***	.957	.608***
Refused answers per completed interview	.820	.553	.331	.221	.068
"Don't Know" or refused answers per completed interview	4.693	3.816***	3.450***	1.168	.676***
Blank answers per completed interview	.543	.846***	.815**	1.967	1.602
Total item nonresponse per completed interview	5.235	4.661	4.265	3.134	2.279***
Length of completed interview in minutes	29.068	30.644*	32.371***	31.390	30.514*

n.a. = Not applicable

a/ All of the significance levels shown in this table are somewhat overstated because adjustments are not made for sample clustering.

b/ Includes 447 potential respondents and 311 completed interviews in three urban sites.

c/ Includes 956 potential respondents and 699 completed interviews in seven urban sites.

d/ For interview completion rates, refusal rates, and hours searching, the five rural sites and two unusual urban sites (one high and one low) are excluded; for the nonanswer and length of interview variables, all 1,185 observations in all twelve payment sites are included.

e/ The overall nonpayment group includes 1,262 observations and is reduced proportionately for the outcome measures that apply only to subsamples (e.g., for youths who completed the interview.

f/ The overall payment group included 3,126 observations and is reduced proportionately for the outcome measures that apply only to subsamples (e.g., for youths who completed the interview).

g/ Observations were excluded if the postal service could not deliver the advance letter (270 youths in the nonpayment group and 540 in the payment group).

* Different from the nonpayment group mean at the 90 percent level of confidence.
** Different from the nonpayment group mean at the 95 percent level of confidence.
*** Different from the nonpayment group mean at the 99 percent level of confidence.