Diagnostics for Redesigning Survey Questionnaires:

Measuring Work in the Current Population Survey[1]

by

Elizabeth Martin, Bureau of the Census

and

Anne E. Polivka, Bureau of Labor Statistics

Shortened title: Diagnostics for Redesigning Questionnaires

Diagnostics for Redesigning Survey Questionnaires: Measuring Work in the Current Population Survey

ABSTRACT

Between 1986 and 1993, a program of questionnaire design and cognitive research was conducted by the Census Bureau and Bureau of Labor Statistics, to improve labor force measurements in the Current Population Survey. As part of the research program, diagnostic measures for systematically testing and evaluating alternative questionnaire versions were developed and applied. This paper reports results of applying two methods, special follow-up probes and hypothetical vignettes, to the measurement of "work" in the CPS. These measures provided both direct and indirect information about problems of respondent comprehension and reporting errors. In this paper we analyze results using these diagnostic measures to evaluate the effect of questionnaire revisions on reporting of work activities, and we assess the consistency and usefulness of the information provided by alternative diagnostic measures for pretesting and selecting questions.

Introduction

How should survey designers go about evaluating the effects of questionnaire revision in a survey designed to produce very precise estimates? This question confronted Government statisticians and methodologists during the redesign of the Current Population Survey. The CPS is the national labor force survey conducted monthly by the Census Bureau for the Bureau of Labor Statistics to produce national and state estimates of employment, unemployment, and other economic indicators. The survey samples 60,000 households per year, and is designed to detect a change of .1 percent in the unemployment rate. The requirement for highly precise measurements increased both the necessity and the difficulty of reliably measuring potential questionnaire effects on the estimates.

This paper evaluates several diagnostic techniques which were used to assess questionnaire alternatives during the redesign process in conbination with large-scale split-sample experiments. The diagnostic techniques, which rely on hypothetical vignettes and direct probing questions, do not require very large samples to provide useful information, and potentially may be applied to attitude surveys as well as factual surveys.

Background

The focus of the research reported here is on the measurement of work in the CPS. In the CPS, respondents who report any work activities (defined as working as little as one hour for remuneration or expected remuneration in a job, business, or on a farm) during the reference week (always the week of the 12th each month) are counted as employed and at work. Prior to 1994, the primary questions used to determine a person's work activities were,

- "What were you [NAME OF OTHER PERSON] doing most of LAST WEEK--

(working, keeping house, going to school) or something else?"[2]

and, if the respondent did not report working,

- "Did you [NAME] do any work at all LAST WEEK, not counting work around the house?"

Respondents may not have known that what should be reported as "work" in the survey did not include volunteer work or school work, especially since reports of such activities seemed to be invited by the initial question. In addition to erroneously reporting such activities as "work," respondents may have failed to report activities which should have been reported, such as unpaid work in a farm or family business, or casual employment for just a few hours, such as mowing lawns or babysitting. The second question's exclusion of "work around the house" may have discouraged reporting of true work activities occurring at home. No explicit questions were asked to determine certain key facts, such as whether anyone in the household had a farm or business and whether any family members did unpaid work for it. Rather, the interviewer was supposed to probe as necessary to identify businesses and farms.

These and other problems with the CPS questionnaire were identified by two national commissions and an in-depth review of the questionnaire early in the research process (President's Committee to Appraise Employment and Unemployment Statistics, 1962; National Commission on Employment and Unemployment Statistics, 1979; Bureau of Labor Statistics, 1986). In 1986, the Bureau of Labor Statistics and the Census Bureau began a joint program of questionnaire design and cognitive research on the CPS, with one goal being to identify and correct errors arising from variability in how respondents and interviewers interpreted key concepts, such as "work," "looking for work," "job," and "business." In order to redesign questions appropriately,

information was needed about how respondents were interpreting these words and phrases, and which types of work situations were most likely to be misunderstood. We needed empirical information on which to base decisions about questionnaire revisions, rather than speculation about how respondents were likely to interpret a question wording. In order to provide diagnostic information, we developed and used several types of measures, including the two to be discussed here, hypothetical vignettes and direct probing questions. (Esposito et al. 1991 discuss the various methods used.)

Vignettes have been used in social research to explore social definitions (for example, what constitutes child abuse; see Rossi and Nock, 1983), but they have not been commonly used for redesigning survey questions. Their use as a methodological tool relies on the fact that vignette responses are sensitive to the context created by prior survey questions, and hence they provide indirect measures of the contextual interpretation of key survey concepts. (Vignettes were used for this purpose in the redesign of the National Crime Survey; see Martin et al., 1986.)

The hypothetical vignettes used in the CPS redesign were intended to measure respondents' interpretations of "work" by testing how consistent their classifications of various marginal or ambiguous activities were with the CPS definition. The vignettes described types of marginal work situations thought to be especially problematic, such as casual labor and work in a family business. In the initial stages of the CPS questionnaire design research conducted during 1988, vignettes were administered immediately after the main CPS interview in a CATI survey. The purpose was to measure interpretations of work in the context of prior CPS questions about work. Thus, context effects were used deliberately as a way to understand respondents' interpretations of a key survey construct. This research demonstrated that respondents held very

diverse interpretations of "work," with only 8 percent of the sample classifying 5 hypothetical work vignettes consistently with the CPS definition (Campanelli, Martin, and Creighton, 1989). Most commonly, respondents interpreted work too broadly (that is, they would report as "work" activities which CPS does not), suggesting the potential for overreporting bias. However, the second most common interpretation was much too restrictive, with respondents ruling out all marginal work activities, including those counted as work by CPS. Older respondents held stricter interpretations of work than younger ones (Martin, Campanelli, and Fay, 1991). The results were consistent with the long-suspected underreporting of teenagers' employment by older proxies. This bias appeared to be the combined effect of a too-narrow interpretation of work by older respondents, and teenagers' tendency to engage in the types of casual employment which older respondents tend not to regard as "work."

Based on the 1988 vignette study and other methodological studies, several questionnaire revisions were made to improve respondents' understanding and reporting of work activities:

- The initial question ("What was [NAME] doing most of last week -- (Working, going to school, keeping house), or something else?") was dropped. This was done to avoid focussing respondents' attention on activities unrelated to paid employment and possibly inviting respondents to report them as "work."

- New questions were added to determine whether anyone in the household had a business or farm and, if so, whether any unpaid work was done in connection with it. This was done to avoid reliance on volunteered information.

- The wording of the work question was revised to emphasize the defining criteria for work which counts as employment in the CPS. The work question was changed from, "Did

[NAME] do any work at all LAST WEEK, not counting work around the house?" to "LAST WEEK, did [NAME] do ANY work for (either) pay (or profit)?" The reference to "profit" was included if there was a business identified in the household.

In a second stage of the research, these and other questionnaire revisions were tested in a 1991 split-sample comparison of the redesigned and old questionnaires in a survey conducted using computer-assisted telephone interviewing (CATI). Households were selected by random digit dialing and were randomly assigned either the old or the revised version of the CPS instrument. They stayed in sample for 4 months, receiving the same CPS questionnaire version throughout their tenure. After completing their final CPS interview, household respondents were administered a debriefing interview which included questions to identify problems of comprehension and the like. CPS interviews were conducted in 3,800-6,000 households per month between July and October 1991.

We posited that the two questionnaire versions created different contexts, that is, different implicit (or explicit) definitions of "work." The introduction and wording of the vignette questions were varied to reinforce the effect of the wording of the work question the respondent had previously received. Respondents who had received the old version were asked, "Earlier I asked you a question about working...Now I want you to tell me how you would answer that question for each of the persons in the following imaginary work situations ...Sam spent 2 hours last week, painting a friend's house and was given 20 dollars. Would you report him as working last week, not counting work around the house?" Respondents who had received the revised work question were asked, "Would you report him as working for pay last week?" We assume the prior CPS questions about work influenced subsequent interpretations of the activities

described in the vignettes. Respondents' classifications of the vignettes, then, provide a measure of how well the context established by prior questioning matches the official definition.

The vignettes required respondents to interpret situations which did not apply to them, and thus could not in themselves support inferences about the effects of misinterpretations on reporting. Probing questions, along the lines of those developed by Belson (1981), were designed to directly measure underreporting in the main survey. Respondents were asked, "In addition to people who have regular jobs, we are also interested in people who may work only a few hours per week. LAST WEEK did (NAME) do any work at all even for as little as one hour?" The probe was asked about the first person listed on the household roster who was reported as not working during the reference week. If valid, the results may permit direct measurement of underreporting bias under alternative versions of a questionnaire.

Households were randomly assigned to be asked either vignettes (one in 10) or direct probing questions (9 in 10) in the final debriefing interview. (Of the 6,006 household respondents 16 or older in their final month in sample, 624 received the vignette series and 5,382 were assigned to be asked probing questions. Of the latter, 2,823 were eligible to be asked the work probe because one or more household members reported no work during the reference week. A total of 2,559 respondents were ineligible for the work probe because there were no nonworkers in the household.)

We use the results of the 1991 survey to address the question of whether the alternative diagnostic measures provided consistent and meaningful information about comprehension errors which affect reporting of work activities. In the first section, we discuss results of the work vignettes, and in the second we examine results of the direct probe for missed employment. We

then assess the diagnostic information they provided in the light of actual reporting differences in the survey. The final section comments more generally on the alternative diagnostic measures.

Results of Hypothetical Work Vignettes

Table 1 presents respondents' classifications of the work vignettes administered in the first respondent debriefing study in 1988 and in the split-sample 1991 CATI test, by questionnaire version. The 1988 results confirmed many of the respondent misinterpretations we and others had suspected (Campanelli, Martin, and Creighton, 1989). Respondents' judgments of whether or not the activities described in the vignettes should be reported as work indicated potential underreporting of activities involved in setting up or conducting a business (J and L), underreporting of casual labor for a few hours (K), and overreporting of volunteer work (O).

Comparisons of the 1988 and 1991 vignette results support two conclusions. The first is that respondents' classifications of vignettes in the 1991 survey were highly sensitive to the context created by the experimental variation in the wording of the work question (cf. second and third columns in Table 1). The wording revision drastically reduced the proportion of respondents who would improperly include volunteer work and activities to help a family member (see vignettes O and N) and generally reduced positive responses to all vignettes which do not involve payment, including work for commission (M) and unpaid work in a family business (J). The first two of these changes are evidence of improvement, but the latter two changes represent a worsening. The questionnaire revision also appears to have made it more likely that respondents correctly include casual paid labor, such as described in vignettes I and K. (One should keep in mind that these results provide information about improvement, or lack thereof, in how respondents classify a particular set of vignettes, but do not permit inferences about

improvements in reporting in CPS.)

Second, except for one item, respondents classified the vignette situations very consistently following the old questionnaire version in both 1988 and 1991 surveys (cf. first and second columns in Table 1). This similarity of results occurs despite sample differences[3] and differences between surveys in the introduction to the vignettes. The one significant difference (a drop from 1988 to 1991 in "yes" responses to vignette L) may be due to a contrast effect caused by the order and content of the vignettes. In the 1988 survey, vignette I immediately followed another which described donating blood for money. Very few respondents classified the latter activity as work (the item was dropped in 1991) and the contrast to selling blood may have made setting up an antique shop seem more like "work."

Respondents' Reasoning about Vignette Situations. In the 1991 survey, respondents were asked to explain why they classified the vignettes as they did. Their reasons confirm that the revised wording of the work question focused their attention on payment. (For example, 65 percent of respondents receiving the new questionnaire, compared to 47 percent receiving the old, gave "he was paid" as a reason for their answer to the painting vignette (Ns = 300 and 295, respectively); similar differences were found for all 7 vignettes.) The old question left respondents freer to interpret work in various ways, and consequently, respondents gave more, and more various, reasons for their vignette responses. Moreover, the old question's exclusion of "work around the house," which was meant to exclude housework, inappropriately introduced location as a factor influencing respondents' classifications. Under the revised question wording, such extraneous criteria were mentioned less frequently or not at all. (For example, 12 percent of respondents receiving the old question, compared to 2 percent receiving the new, ruled out setting up an

antique shop in a back room because it was "work around the house"; Ns = 292 and 292.)

Respondents' narrower focus on pay accounts for their correctly including marginal work activities, but also is consistent with some respondents' exclusive reliance on present payment. This overly simple rule led to more incorrect exclusions of legitimate work activities not yet yielding pay or profit, such as described in M and J. Thus, the revised wording did not completely eliminate all errors, but did create a narrower economic interpretation of work which should lead to more uniform reporting in the CPS.

The structure of responses to work vignettes. Another way of using vignettes as a methodological tool for redesigning survey questions would be to combine "yes" responses to the entire set into a scale. The assumption would be that each item measures an aspect of respondents' interpretations of work, and that summing responses across vignettes yields a global measure of how broad or narrow their interpretations are. Such a measure could be used to assess whether a questionnaire revision globally broadens a construct or narrows it. Alternatively, one might score each response as "correct" or "incorrect" in terms of a survey definition, and sum "correct" responses to produce a scale of the congruence of respondents' interpretations with the definition. Such a scale could be used to assess the effects of a questionnaire revision on the consistency of respondents' interpretations with key survey definitions.

As an initial examination of the global effects of questionnaire revision on interpretations of work, two summary indexes were computed based on responses to the 7 vignettes: probability of a correct response, and probability of a yes response. (For each respondent, the number of correct--or "yes"-- responses was divided by the number of vignettes to which responses were given.) The probability of responding correctly in terms of the CPS definition is

essentially unchanged (.67 in the old and .69 in the revised version). However, the average probability of "yes" is significantly lower in the revised questionnaire (.53 in the old compared to .42 in the revision; the standard error of the difference is .018). Thus, the questionnaire revision appears to have created a more restricted frame of reference; fewer activities are likely to be included as "work."

Are responses to vignettes better explained by assuming that respondents' interpretations of work vary in accuracy, or in their inclusiveness? An earlier phase of the CPS redesign research addressed this question by analyzing the structure of responses to work vignettes in the 1988 respondent debriefing study. Martin, Campanelli, and Fay (1991) analyzed various log linear models, including measurement models developed by Rasch (1960/1980), and found that respondents varied in the degree to which they held broad versus narrow interpretations of "work," but there was no evidence of an underlying tendency to answer correctly, in terms of the CPS definition. Their results were interpreted as evidence of a latent tendency to hold inclusive or restrictive views of work which accounted for responses to vignette items. Additionally, there were significant associations between particular pairs of vignettes, suggesting that respondents may apply heuristics or rules to classify situations with similar features, beyond a global tendency to interpret work broadly or narrowly.

To examine the stability of vignette response structure, the earlier analysis was replicated using 1991 data. The model which described the 1988 data was fitted to data for the half sample which received the old questionnaire in 1991, using vignette items common to both surveys. The same model provided a good fit to both datasets, and parameter estimates were closely comparable.[4] Thus, the vignette response structure is stable in independent surveys in which

sample selection and time of administration varied, but in which the prior survey questions and the vignettes themselves were repeated. This result bolsters our interpretation of vignette responses as a measure of the contextual effect of prior questions.

There are several hypotheses of interest about possible effects of the questionnaire revision on interpretations of work. One is that the questionnaire revision had no effects on interpretations of work activities (this hypothesis is already in doubt based on inspection of Table 1). A second is that the questionnaire revision affected the interpretations of particular situations. (This hypothesis would be supported if the data were fitted by a log linear model including only pairwise associations between questionnaire version and vignette responses. That model would imply that classifications of individual vignettes were significantly affected by the questionnaire revision, but that the consistency--or lack thereof--of respondents' classifications of different vignettes was not significantly changed.) A third hypothesis is that the questionnaire revision, by making certain criteria salient, strengthened associations between responses to similar vignettes. (This hypothesis would be supported by the presence in the model of three-way or higher order interactions involving questionnaire version and 2 or more vignettes. For example, since the revised question explicitly mentions "work <u>for pay</u>", one might expect that the vignettes mentioning pay would become more strongly associated. This could happen if the wording revision led respondents to apply this criterion more deliberately and consistently in classifying vignettes.) A fourth hypothesis is that response patterns are explained by a latent tendency to interpret work broadly or narrowly regardless of the particular vignette. This hypothesis is consistent with the Rasch measurement model, and would imply that the effects of the questionnaire revision arose from a latent or global tendency to be more or less inclusive of

different work activities, rather than from changed interpretations of particular situations.[5]

These hypotheses about the effects of the questionnaire revision on interpretations of work cannot be tested by examining bivariate relationships, such as those presented in Table 1. To examine the hypotheses, we fit three log linear and Rasch measurement models to a 6-way cross-classification of responses to 5 vignettes by questionnaire version, scoring the items "yes/no". Two vignettes to which very few people responded "yes" in the revised questionnaire (O and N in Table 1) were dropped because of the large number of zero cells in the 8-way cross-classification.

Model 1, the model of statistical independence, assumes that responses to each of the five vignettes are independent of one another and do not vary by questionnaire version. This model fits poorly ($X^2 = 332.25$, d.f. = 57, $p < .0001$), indicating that there are significant associations among responses to vignettes, and/or between vignette responses and questionnaire version.

Model 2 assumes associations among all pairs of vignettes, and between questionnaire version and response to each vignette, but no higher-order interactions. This model fits the data well, with $X^2 = 44.11$ on 42 degrees of freedom, and probability of .45. None of the possible 3-way interactions were statistically significant, which implies that the association between responses to any pair of vignettes does not depend on responses to a third. It also means that questionnaire version did not significantly affect associations between items. Thus, although the respondents' reasons for their vignette responses suggested they were more focussed on the criterion of payment, results of this analysis suggest the questionnaire revision did not significantly increase associations between vignettes which mentioned pay.

Model 3 is the Rasch model, which assumes that the associations among response

variables are entirely accounted for by an underlying dimension represented by Rasch parameters, expressing heterogeneity among individuals in the latent propensity to view different situations as "work." This model posits that once variation among individuals in the underlying trait is explicitly accounted for by including the Rasch parameters in the model, then associations among individual response variables should vanish. Although the fit of the Rasch model is better than statistical independence, it is quite poor ($X^2 = 228.75$, d.f. = 48, $p < .0001$). It is possible to add pairwise associations to this model, as was done in the analysis of the 1988 data, and obtain a good fit to the data. However, the most parsimonious models with acceptable fit do not include the Rasch parameters.

In sum: Evidence shows that the responses to individual vignettes were highly sensitive to the context created by prior questions, and pointed to specific areas of misunderstanding which required questionnaire revision. The response structure of the vignette scale appears stable in two independent surveys in 1988 and 1991, supporting our use of the scale as a measure of context-specific interpretations of work. The questionnaire revision resulted in a more restricted frame of reference in respondents' interpretations of work. The reasons given for their vignette responses suggested they focussed more exclusively on "pay." The results of modelling the 1991 vignette data suggest that the questionnaire revision had straightforward effects on interpretations of work: respondents' understandings of particular types of situations were modified. The 1991 data for both questionnaire versions are well-described without assuming that the questionnaire revision globally broadened or narrowed respondents' interpretations of work. As was found in the 1988 analysis, there is evidence of pairwise associations indicating respondents classified similar types of situations consistently. These results are consistent with the second hypothesis

discussed above, but do not support the first, third, or fourth hypotheses. The modelling results are also consistent with Martin, Campanelli, and Fay's (1991) analysis of the 1988 data.

Results of the Direct Probe for Missed Work Activities

As noted above, household respondents received a debriefing interview their final month in sample, and respondents in 9 out of 10 households were probed about the work activities of the first person over 15 years old listed on the household roster for whom no work activities had been reported in the main interview. (The probe was, "LAST WEEK did NAME do any work at all, even for as little as one hour?") About 2 percent of persons for whom no work was reported in the main CPS interview reported bona fide work activities when probed.[6] The difference between the two versions of the CPS questionnaire was nonsignificant (2.87 and 2.31 percent of those probed in debriefing mentioned work activities they had failed to report in the old and new questionnaire, respectively, based on samples of 1,465 and 1,253).[7]

Although this measure provides no statistically significant evidence of questionnaire effects on overall reporting of marginal work activities, there is evidence of improvement for a particular type of activity targetted by questionnaire revisions: work in connection with a family business or farm. In the old CPS, missed work was over twice as common in households in which there was a business, as shown in Table 2. The new questionnaire added questions about the presence of a business in the household and work in connection with the business. As a result, in the new questionnaire, work is only slightly less likely to be underreported in business households than in others. (However, the improvement in reporting for business households, indicated by a decline from 7.0 to 3.0 percent of persons whose work activities were missed, is statistically

insignificant, due to small sample sizes.)

Tables 3-4 show the relationships between missed work and the age and gender of the reference person, for each version of the questionnaire. These tables reveal striking differences among demographic groups in the proportion of persons whose work activities were not identified in the main CPS interview. In general, patterns of missed work are quite similar under both versions of the questionnaire. Table 3a shows that the proportion of persons who report work activities in response to the debriefing probe declines monotonically with age. The relationship between missed work and age was statistically significant in the old questionnaire but not the new; missed work is lower in the new questionnaire, but not significantly so.

Table 3b shows estimates of the effect of unreported work activities on estimates of the proportion of each age group which was employed and at work. To produce an aggregate national estimate, responses to the work probe were weighted by the number of persons in each household who were eligible for the probe.

The proportion of workers which the main survey failed to identify are dramatically higher among both young (16-19) and old persons (post retirement, 65 and older) than persons in their middle years, for whom rates are quite low. Results are quite similar for both questionnaire versions. One of the aims of the redesigned questionnaire was to reduce the amount of missed employment among youth, which has long been hypothesized as a source of bias in CPS (National Commission on Employment and Unemployment Statistics, 1979). These results provide evidence of that bias, which was reduced by nearly a third in the new questionnaire, from 3.8 to 2.6 (this difference is not significant, based on 230 and 213 unweighted cases, respectively). The proportion of workers of retirement age who were not identified in the main survey declined by

over 20 percent from 11.9 to 9.4 percent.[8] Since the new questionnaire features a greatly curtailed set of questions about work activities for retired persons, it is somewhat reassuring to find that the reduced number of questions did not lead a larger proportion of post-retirement workers to be misclassified as not working. However, these data suggest high rates of misclassification error for this group for both questionnaire versions, and indicate an area where further improvements appear needed.

Table 4a shows the association of gender with missed work, with a larger proportion of men reporting missed work activities in response to the debriefing probe than women; this difference is statistically significant only in the new questionnaire and is due primarily to a decrease in missed work for women. The effect on estimates of the work force was marginally significant ($p<.09$), but only for the old questionnaire, as shown in Table 4b. A slight gender bias which resulted in underreporting of the number of female workers relative to men in the old questionnaire was eliminated, and slightly reversed, in the new questionnaire.

In sum, the results of the direct work probe imply the proportion of persons at work will remain the same, or perhaps increase slightly, using the new questionnaire. Underreporting of work activities in households with businesses should decline. The relationship between underreporting and age should remain approximately constant, while a small gender bias has been eliminated.

Implications of the Diagnostic Measures for Reporting Differences

The diagnostic information provided by the vignettes suggested that the questionnaire revision had both positive and negative effects in terms of creating a frame of reference consistent with the intended meaning of work. On one hand, the narrower focus of the revised work question should

be beneficial in reducing reporting of non-work activities, such as volunteer work. (Fewer such mentions may explain interviewers' reports that they did not have to probe as much with the revised work question; see Polivka and Rothgeb, 1992.) On the other hand, the results could also imply reduced reporting work for commission (see M) and unpaid work in a family business (see J). However, since the revised questionnaire eliminates reliance upon volunteered reports of unpaid work in a family business through the inclusion of direct questions, the importance of the latter finding is reduced. Adding direct questions targeted to these activities should improve reporting, regardless of whether respondents would report them in response to a general question about work.

As noted earlier, the vignettes describe hypothetical situations. We assume that there is some correspondence between how respondents classify a vignette and how they would report a similar situation which applied to them or their household members in the survey. Although we cannot test this assumption directly, it is of interest to examine the consistency of the questionnaire differences in reporting of actual work activities with the differences implied by the results of the vignettes.

The results of the analyses of the vignettes support the following expectations:

(1) The proportion of persons reporting work activities during the reference week should be slightly reduced in the revised questionnaire due to the narrower frame of reference, except that the addition of direct questions about unpaid work in a family business or farm should lead to better reporting of those activities.

(2) The questionnaire revision should lead to better reporting of casual employment, such as work for a few hours or by teenagers or students.

(3) The questionnaire revision may lead to reduced reporting of work for commission or work compensated by other than "pay."

The results of the direct probe do not support a strong prediction about possible effects of the questionnaire revision, but suggest that missed work should be slightly reduced in the new questionnaire, especially in households in which there is a business. The relationship between underreporting and age should remain constant.

Evidence from the 1991 split-sample CATI test shows a very small overall difference between questionnaire versions in the proportion of persons reported as working during the reference week: 57.7 and 58.2 percent in the old and redesigned questionnaires, respectively, based on samples of 16,175 and 15,609 persons (as opposed to households). However, the revised question wording results in a significant increase, from 40.8 to 45.5 percent, in the proportion of persons 16-19 years old reported as working during the reference week, as shown in Table 5.[9] There are no significant questionnaire differences for other age groups. Using a logit model, the interaction effect between age, questionnaire version, and proportion reported as working is statistically significant using 2 categories of age (19 years old and younger, 20 and older) ($X^2 = 6.65$, df=1, $p<.01$). Thus, the revised question wording led to increased reporting of work activities for teenagers, but not older respondents. The proportion of currently-enrolled students identified as also working is slightly higher in the revised questionnaire (45.4 and 48.8 percent in the old and redesigned questionnaires, respectively, with N=930 and 1,045) but the difference is not statistically significant.

The revised questionnaire elicits more reports of work in a person's own or a family business or farm. The proportion of workers classified as self-employed or as unpaid workers in a

family business rose significantly from 12.9 percent in the old questionnaire to 14.8 percent in the redesigned questionnaire (N=3,384 and 3,244, respectively; X^2=4.6, df=1, p<.032).

The revised questionnaire also elicits more reports of work activities involving relatively few hours. The proportion of workers reporting 19 or fewer hours of work during the reference week is 10.0 percent in the revised questionnaire compared to 8.8 in the old (N=9,028 and 8,652, respectively; X^2=8.5, df=1, p<.003). The difference is small but consistent with the expectation that the revised questionnaire picks up more casual labor.

Finally, contrary to the expectation based on vignette analysis, there is scant evidence that the revised questionnaire does a worse job of identifying work on commission. The proportion of persons at work during the reference week who reported in the debriefing interview that they usually receive commissions is 11.4 and 10.4 percent for the old and revised questionnaires (N=518 and 568, respectively; the difference is not significant).

Conclusions

The diagnostic measures developed and refined in the CPS questionnaire redesign effort appear to yield consistent and useful information about the nature and extent of response biases present in the data. In the research reported here, the different methods for evaluating the effects of question wording changes provided different kinds of information to questionnaire designers. The CPS is intended to produce very precise measures of labor force characteristics, but in many cases, there is no external evidence with which to verify whether a respondent was "really" working or looking for work or not, so the evaluation of the effect of questionnaire revisions must depend on internal evidence about the quality and consistency of the data. Moreover, split-sample experiments have somewhat limited utility in this context, because they must be so enormous to

detect effects of question wording changes upon labor force estimates. In this context, the indirect measures used in this research provided information about the performance of the redesigned instrument which supplemented information from split-sample experiments. We believe these methods are applicable to a variety of survey situations. The vignettes proved useful for identifying problems of comprehension and question wording, and their sensitivity to changes in the wording of prior questions implies that wording effects can be detected with relatively small sample sizes. The vignettes appear to yield comparable results when they are replicated in independent surveys following the same main survey instrument. This suggests that sets of vignettes may be relatively reliable measures of the contextual meaning of key survey constructs. Their sensitivity and robustness recommend them as tools for redesigning questionnaires when the designer needs to know how well respondents' interpretations of key words and phrases matches the intended meaning. The vignette technique seems especially applicable when a survey requires respondents to make implicit or explicit judgments about the scope of complex phenomena which are the subject of questioning, and could usefully be applied to design and test surveys of attitudes as well as factual topics. It should be noted, however, that their sensitivity to wording effects may be a drawback in some cases, because they may be sensitive to unintended as well as intended context effects. An example was an apparent contrast effect caused by a prior vignette included in the 1988 survey but not in 1991.

Fitting alternative measurement models to a vignette scale can be used to examine hypotheses about the extent to which a questionnaire defines meanings which are global or particular, whether the respondent is using rules or heuristics to judge situations, and the types of situations which are seen as similar by respondents. Analysis of the reasons respondents gave for

their vignette responses, suggested that the intended specificity was not achieved under the old questionnaire. When the goal is specificity, the failure of the Rasch model to adequately describe vignette data might be taken as evidence of improvement. However, in other applications, the questionnaire designer may intend to measure a global construct, abstracted or generalized from particular situations. In such cases, it would be desirable to find that the Rasch model fits, and that respondents adopt global rather than specific rules for classifying different situations.

The direct probe for missed work provides new information about the underreporting of work activities, and shows promise as a way to directly estimate underreports. Review of the responses to the work probe showed that a large portion represented bona fide work activities, indicating that the probe yields largely valid data. Bias estimates, such as the work probe, may provide measures, lacking until now, by which to judge future redesign efforts in CPS. Direct probes provide useful information about the types of activities which are not reported in the main survey, and about the relationship between underreports and respondent characteristics. The latter is not available from analysis of hypothetical vignettes or other sources. In this study, results of a probe for missed work provided the first direct evidence of the long-suspected underreporting of teenage employment.

More generally, we suggest that two methodological developments can advance the art (and science) of questionnaire design. The first is the design of methods which permit quantitative measurements as well as qualitative assessments of the degree to which a survey questions meet the objectives of their designer. The vignette technique makes deliberate use of context effects as a tool for designing questionnaires: by manipulating the effect of different question wordings on vignette responses, the questionnaire designer can measure the degree to

which a given wording is interpreted as intended by respondents. The direct probing technique provides a potential measure of underreporting bias. These tools provide new types of information, which in the past have not been available to questionnaire designers. Both techniques need further development and exploration before their usefulness can be judged. The second necessary advancement is to fit measurement models of response structure as part of the questionnaire design process, in order to better understand the conceptual underpinnings of respondents' answers to survey questions. The point is, direct probes, sets of vignettes, and the measurement models applied to analyze responses to them, can provide new and sophisticated information to the questionnaire designer about whether survey design objectives are achieved by a questionnaire.

References

Belson, W. A. (1981.) The Design and Understanding of Survey Questions. London: Gower.

Bureau of Labor Statistics. (1986.) Report of the BLS-Census Bureau Questionnaire Design Task Force. Washington, DC: BLS.

Campanelli, P. C., E. A. Martin, and K. P. Creighton. (1989.) Respondents' Understanding of Labor Force Concepts: Insights from Debriefing Studies. Proceedings of the Fifth Annual Research Conference, pp. 361-374. Washington, DC: Bureau of the Census.

Duncan, O. D. (1984.) Rasch measurement in survey research: further examples and discussion. In Surveying Subjective Phenomena, Vol. 2, edited by C. F. Turner and E. A. Martin. New York: Russell Sage.

Esposito, J. L., P. C. Campanelli, J. Rothgeb, and A. E. Polivka. (1991.) Determining which Questions Are Best: Methodologies for Evaluating Survey Questions. Proceedings of the American Statistical Association (Survey Research Methods Section).

Goodman, L. A. (1978.) Analyzing Qualitative/Categorical Data. Cambridge, MA: MIT Press.

Martin, E. A., P. C. Campanelli, and R. E. Fay. (1991.) An Application of Rasch Analysis to Questionnaire Design: Using Vignettes to Study the Meaning of "Work" in the Current Population Survey. The Statistician 40: 265-276.

Martin, E. A., R. Groves, J. Mattlin, and C. Miller. (1986) Report on the Development of Alternative Screening Procedures for the National Crime Survey. Washington DC: Bureau of Social Science Research.

National Commission on Employment and Unemployment Statistics. (1979.) Counting the Labor Force. Washington DC: GPO.

Polivka, A. E. and J. M. Rothgeb. (1992.) An Illustration of the Methods Used to Redesign the Current Population Survey Questionnaire. Presentation for the annual meeting of the American Association for Public Opinion Research, May 1992.

President's Committee to Appraise Employment and Unemployment Statistics. (1962.) Measuring Employment and Unemployment. Washington DC: GPO.

Rasch, G. (1960/1980) Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago Press.

Endnotes

1.This paper reports the results of research undertaken by Census Bureau and Bureau of Labor Statistics staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or the Bureau of Labor Statistics. The authors thank Douglas Bond, Jennifer Hess, William Nicholls, Paul Siegel, and Colm O'Muircheartaigh for helpful comments.

2.The household respondent answers the labor force questions for each individual in the household 16 and older. The interviewer was instructed to select the phrase in parentheses which seemed most appropriate to ask about a particular person; for example, "keeping house or something else" would be asked about persons who appeared to be homemakers.

3.The 1988 survey was a probability sample of 30 multi-interviewer PSUs selected from 109 metropolitan areas, while the 1991 sample was selected by random digit dialling. Both surveys were conducted by CATI.

4.The 4-way cross-classification of 1988 responses to vignettes J, K, L, and O was well-described by a model fitting the one-way marginals of each vignette, the pairwise associations between vignettes J and L, J and O, and K and O, as well as the Rasch parameters ($X^2 = 5.76$, d.f. = 5, $p < .3$). The same model also provided an adequate fit to 1991 data for the old questionnaire ($X^2 = 4.3$, d.f. = 5, $p < .4$).

5. The Rasch measurement model treats each response as the product of two parameters: one unique to the item, and the other unique to the person. When items are scored "yes/no", the person parameter represents an individual's latent tendency to interpret the concept of work inclusively or restrictively. It is assumed there is an underlying dimension along which individuals are arrayed according to the inclusiveness of their interpretations of "work." Respondents

answering "yes" to many of the vignettes would be toward the inclusive end of the underlying interpretation-of-work dimension; those rejecting many of the vignette situations as work are toward the restrictive end of the latent dimension. The item parameter represents how difficult or easy an item is; in other words, how congruent the activity described in the vignette is with respondents' underlying concept of "work." (For instance, Table 1 suggests it is easier for most respondents to view getting paid to tend bar as "work", than babysitting a grandson.) (See Goodman, 1978; Duncan, 1984, for detailed discussions of the application of Rasch measurement models to categorical data.)

6.Positive responses to the probe were followed up with, "What kind of work did NAME do?" and "Did NAME get paid for the work?" Review of the verbatim responses suggest that 20 of 96 positive responses did not describe "work" as defined by CPS. These are excluded from figures reported in the paper. Other, valid responses describe primarily casual labor (babysitting, mowing lawns or yardwork), odd jobs or on-call work (security jobs, cleaning houses for pay, parking cars, construction and repair), farmwork or unpaid work on a family business, and work for pay which somehow went unreported, perhaps because it did not seem like "work" or a "job" (including teaching and tutoring, therapy, art, music and craftwork).

7.The reader should keep in mind that differences which would be regarded as trivial in other analyses often are substantively and statistically significant for labor force estimates. Recalling that the survey is designed to detect a change of .1 percent in the unemployment rate, the reader can appreciate that apparently small differences in employment are important.

8.The difference in the results in Table 3a and 3b result largely from changes in the denominators. Table 3a is based on those identified as not working in the main survey; Table 3b is based on all

workers, whether identified in the main survey or by the debriefing probe. Thus, the number of persons 65 and older who failed to report their work activities in the main survey is tiny as a fraction of nonworkers, but quite substantial as a fraction of workers of that age. The contrasting results in Tables 3a and b show that a small rate of error can substantially affect the estimates for small groups, such as workers 65 and older.

9.This difference is much larger than would be expected based on Table 3b, which predicts an increase from 40.8 to 41.6 in the percent of 16 to 19 year olds who were working--that is, an increase of 3.8 - 2.6 = 1.2 percent.

Table 1

Percent Reporting Vignettes as "Work" following Alternative Questionnaires in 2 Surveys

			Percent classifying vignette as "work"		
	Vignettes describing work situations	Survey year	1988	1991	
	"Earlier I asked you a question about working. Now I want you to tell me how you would answer that question for each of the persons in the following imaginary work situations...."[a]	Question-naire version	Old	Old	New
I	Bill attended his college classes and got paid to tend bar for a fraternity party one night last week. (1991, old q'aire): Would you report him as working last week, not counting work around the house? (1991, revised q'aire): Would you report him as working for pay (or profit) last week?	n.a.	78		85
J	Last week, Amy spent 20 hours at home doing the accounting for her husband's business. She did not receive a paycheck. (In 1988): Do you think she should be reported as WORKING last week?	50	46		29
K	Sam spent 2 hours last week, painting a friend's house and was given 20 dollars.	64	61		71
L	Last week, Sarah cleaned and painted the back room of her house in preparation for setting up an antique shop there.	59	47		42

M	Cathy works as a real estate agent for commissions. Last week she showed houses but didn't sign any contracts.	n.a.	89	61
N	Fred helped his daughter out by taking care of his grandson two days last week while the boy's mother worked.	n.a.	13	2
O	Last week, Susan put in 20 hours of volunteer service at a local hospital.	38	36	4
	Total N asked work vignettes	1,980	305	319

[a]Note: In 1988, the introduction was, "I asked you a question about WORKING last week. Now, I'm going to read you a list of examples. After each example, please tell me whether or not the person should be reported as WORKING last week." According to CPS criteria, correct answers are "Yes" to vignettes I, J, K, L, and M, and "No" to vignettes N and O. Vignettes were asked in the order I, O, J, K, L, N, M in the 1991 survey. The order was O, J, K, L in 1988. Missing data are excluded from the calculations. (Item nonresponse rates were between 4 and 7 percent for the vignettes.)

Table 2

Presence of a business in the household, and missed work[a]

	Old questionnaire		Revised questionnaire	
Is there a business or farm in the household?	Percent of Rs reporting missed work when probed	N	Percent of Rs reporting missed work when probed	N
Yes	7.02	114	3.03	132
No business	2.52	1351	2.23	1121
Total	2.87	1465	2.31	1253
	X^2=7.65, df=1, p<.006		X^2=.33, df=1, n.s.	

[a]Missing data for the work probe are excluded from calculations.

Table 5

Proportion of Persons Reported as Working Last Week,

by Questionnaire Version and Age

	Old Questionnaire		Revised Questionnaire	
Age	Percent	N	Percent	N
16-19 years old	40.8	1,594	45.5	1,465
20-29	72.3	2,772	71.6	2,804
30-39	73.8	3,507	74.5	3,446
40-64	65.4	5,836	65.8	5,429
65 and older	11.3	2,466	10.7	2,465
Total	57.7	16,175	58.2	15,609

Table 3

Relationship between Age and Missed Work, by Questionnaire Version

Age	a. Percent of persons reporting missed work when probed			
	Old questionnaire		Revised questionnaire	
	Percent	N	Percent	N
16-19 years old	6.7	135	5.2	116
20-29	4.2	165	3.4	175
30-39	4.0	201	2.7	187
40-64	2.1	427	1.5	342
65 and older	1.7	537	1.6	433
	X^2=12.6, df=4, p<.02		X^2=7.3, df=4, n.s.	

Age	b. Estimated percent of workers unreported in the main survey[a]			
	Old questionnaire		Revised questionnaire	
	Percent	N	Percent	N
16-19 years old	3.8	265	2.6	227
20-29	.8	834	.7	839
30-39	.7	1092	.6	1052
40-64	.9	1663	.7	1495
65 and older	11.9	118	9.4	96
	X^2=71.0, df=4, p<.0001		X^2=62.7, df=4, p<.0001	

[a]Estimates were created by weighting each response to the work probe by the number of persons in a household who were eligible for the probe. The numerator is the weighted number of persons identified as working in response to the probe. The denominator is the total number of workers identified either in the main interview or in response to the work probe. Chi-square values are calculated on unweighted data.

Table 4

Relationship between Gender and Missed Work, by Questionnaire Version

Gender	a. Percent of persons reporting missed work when probed			
	Old questionnaire		Revised questionnaire	
	Percent	N	Percent	N
Male	3.2	524	3.9	413
Female	2.7	941	1.6	840
	X^2=.4, df=1, n.s.		X^2=6.6, df=1, p<.01	

Gender	b. Estimated percent of workers unreported in the main survey			
	Old questionnaire		Revised questionnaire	
	Percent	N	Percent	N
Male	1.2	2155	1.1	2038
Female	1.5	1817	.9	1671
	X^2=2.9, df=1, p<.09		X^2=.1, df=1, n.s.	