RESEARCH REPORT SERIES
(Survey Methodology #2007-30)

An Application of Rasch Analysis to Questionnaire Design: Using Vignettes to Study the Meaning of "Work" in the Current Population Survey

Elizabeth Martin
Pamela C. Campanelli
Robert E. Fay

Director's Office
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 23, 2007

An Application of Rasch Analysis to Questionnaire Design: Using Vignettes to Study the Meaning of "Work" in the Current Population Survey

Paper presented at the
International Conference on Survey Design, Methodology, and Analysis
University of Essex, Colchester, England
July 4-7, 1990

Elizabeth A. Martin, Pamela C. Campanelli, and Robert E. Fay
U.S. Bureau of the Census, Washington DC 20233 U.S.A.

1 Introduction

A basic premise of the survey method as currently practiced is that standardization is necessary to yield comparable data among respondents. Questions are to be read verbatim in a set order by interviewers, who commonly are instructed not to interpret or paraphrase questions or define terms for respondents. The assumption is that questions worded the same way will be understood the same way. It is rare for survey researchers to test this premise, however, and critics have challenged the assumption that standardized question wordings yield standardized question meanings (Mishler, 1986; Suchman and Jordan, 1990).

In this paper we present evidence concerning respondent comprehension of key questions and concepts in the Current Population Survey (CPS), the national household survey used to obtain monthly statistics on employment, unemployment, and labor force participation for the civilian U. S. population aged 16 and older. The research reported here is part of a collaborative,

interagency program of research and development by the Bureau of Labor Statistics and the Bureau of the Census to redesign the CPS. A major objective is to redesign the questionnaire to improve respondent comprehension of questions and concepts. Currently, the CPS does not define labor force concepts for respondents but instead relies on their understanding of "work," "job," "business," and other terms. Although these are familiar, everyday words, it is not clear that shared meanings exist which match the official CPS definition (see Rothgeb, 1982; Bailar and Rothwell, 1984; Bureau of Labor Statistics, 1986; Martin, 1987; Schwarz, 1987; Campanelli, Martin, and Creighton, 1989 for discussions of problems of ambiguity of labor force concepts). We report the results of a new method for testing the meaning of survey questions, based on the development of scales composed of vignettes describing hypothetical situations pertaining to key survey concepts.

2 The Measurement of "Work"

Each month, the CPS counts as employed all persons who were working during the week of the 12th of that month. The official definition includes work for wages or salary, at piece rates, on commission, for tips, or for payment "in kind." Work for the purpose of earning a profit or fees on a family business or farm is included, whether or not the individual had earnings. Volunteer work, school work, and housework are excluded. The determination that a person is working is based, in part, on a respondent's answer to the following question.

"Did (NAME) do any work at all LAST WEEK, not counting work around the house?"

Interviewers are instructed to ask about unpaid work if there is a farm or business operator in the household.

Two types of comprehension problems may affect the validity of respondents' reports of

whether they worked or not. First, respondents may be reporting in terms of an interpretation of "work" that does not agree with the CPS definition. Past research (Belson, 1981) suggests that respondents' interpretations may disagree with a survey definition by either being too broad or too narrow, in terms of the technical meaning intended in the survey. In either case, the survey designer is not measuring what he or she intends to measure. If respondents all interpret a question or term the same way, however, at least the data are comparable across respondents.

A second type of comprehension problem occurs if respondents disagree among themselves on interpretations of survey questions. Lack of consensus on the meaning of key survey terms can seriously affect the comparability of data, particularly if the differences of interpretation are large and coincide with boundaries between subgroups defined by culture, race, age, or other characteristics. An example, from a recent ethnographic study of the Haitian community in Miami, illustrates the point. According to Stepick and Stepick (1990:44-45):

> Obtaining reliable and valid employment data on Haitians is virtually impossible. In the ethnographic research and in the first random survey, more than 50 percent of the Haitian population claimed to be unemployed, when in fact many were engaged in temporary, part-time, or informal sector activities.... These range from sewing a dress now and then for a few dollars to an informal bus transportation operation that generates nearly a thousand dollars a week....[F]undamental is a conceptual difference in the seemingly straightforward question: "Are you working?".... The only condition that elicits a "yes" ... is when the individual has a full time, permanent position.

The Haitians may represent an extreme case, but the implication of these observations is that employment for the Haitians in the United States, and perhaps for other groups as well, could be under-reported in a labor force survey, relative to groups who hold a less restrictive interpretation of work.

3 Methodology

In order to test respondents' understanding of key CPS concepts, such as work, we developed brief vignettes which described hypothetical work situations, such as the following: "Sam spent 2 hours last week painting a friend's house and was given 20 dollars." The vignettes were presented to respondents who had just completed the CPS interview, with the following instruction:

> "I asked you a question about WORKING last week. Now, I'm going to read a list of examples. After each example, please tell me whether or not you think the person should be reported as WORKING last week."

All of the vignettes pertained to marginal or problematic classification situations (see Table 1). By asking respondents to interpret hypothetical but realistic work situations, we hoped to learn more about how the social definition of work varies among respondents, and how consistent respondents' interpretations are with the technical definition of work in the CPS.

Table 1. Respondents' understanding of the concept of work

I asked you a question about WORKING last week. Now, I'm going to read a list of examples. After each example, please tell me whether or not you think the person should be reported as WORKING last week.

CPS definition		Vignette	Yes	*n*
(WORK)	(I)	Last week, Amy spent 20 hours at home doing the accounting for her husband's business. She did not receive a paycheck. (Do you think she should be reported as WORKING last week?)	50%	1977
(WORK)	(J)	Sam spent 2 hours last week painting a friend's house and was given 20 dollars. (Do you think...)	64%	1976
(WORK)	(K)	Last week, Sarah cleaned and painted the back room of her house in preparation for setting up an antique shop there.	59%	1949
(NON-WORK)	(L)	Last week, Susan only did volunteer work at a local hospital.	38%	1973
(NON-WORK)	(M)	Laura earned some money last week by donating blood.	14%	1980

Note: items were asked in the order L, I, J, M, K.

Vignettes have been fruitfully used in social research to investigate variations across the population in social definitions and normative judgments, such as what constitutes sexual harassment or child abuse, the fairness of earnings and the severity of crimes (see Rossi and Nock, 1983; Wolfgang et al., 1985). However, the use of vignettes as a methodological tool to identify problems of respondent comprehension in surveys is relatively new, and was first developed for the redesign of the National Crime Survey (Martin et al., 1986; Biderman et al., 1986).

The response rates for the CPS interview were high (greater than 97%), yielding 2,298 respondents who were eligible to be asked the vignettes and other debriefing questions. Of these,

98% initially agreed to participate but only 92% completed the debriefing interview. Interviews were conducted in July through December of 1988 using the Census Bureau's computer-assisted telephone interviewing (CATI) facility. This made it possible to tailor which debriefing questions respondents were asked according to which questions they had been asked in the main interview. On any given question, no more than 1% of respondents refused to answer and (with two exceptions) no more than 3% said they did not know the answer to a question. Because of an error in CATI programming, however, about 100 respondents who should have been asked the work vignettes were not. (This error affected data only from the first month of the study.) Due to the nature of the sample[1], the data cannot be considered nationally representative. Analyses are based on unweighted and unadjusted data.

As shown in Table 1, the five vignettes pertaining to work include volunteer work and donating blood for money (see L and M), which the CPS does not classify as work, and unpaid work in a family business, casual labor painting a friend's house for $20, and setting up an antique shop in one's home (see I, J, and K), which the CPS does classify as work. For each vignette, most respondents correctly classified the situation as work, or not, as defined by the CPS. However, only 8% of respondents gave responses to all five vignettes which agreed with the CPS definition, as shown in Table 2. The most common response pattern was too inclusive, including as "work" volunteer work, which the CPS does not. The next most frequent response pattern was far too strict, excluding all marginal or unconventional work activities.

---

[1]The data from the Respondent Debriefing Study represent part of a probability sample concentrated in 30 multi-interviewer metropolitan areas drawn from a frame of 109 metropolitan areas. Some households were excluded from the CATI interviewing and subsequently the debriefing study, including those without a telephone and those judged as difficult to enumerate for reasons such as a language barrier or borderline refusal. The effect of the exclusion of these cases from the Respondent Debriefing Study is unknown.

Table 2. Patterns of response to five hypothetical work activities

Activities reported as "work"	Percentage who would report
(1) All activities except blood donation (I, J, K, L)	15
(2) No activities	12
(3) Painting for $20, only (J)	10
(4) Painting for $20, setting up an antique shop (J,K)	9
(5) Accounting, painting for $20, setting up antique shop (I, J, K, official CPS definition)	8
Other patterns of response	46
Total who answered all five items	100
N	1771

As noted earlier, respondents' failure to mention *bona fide* work activities is a source of under-reporting bias which is especially serious if it is correlated with demographic characteristics. If one demographic group holds a restrictive definition and therefore fails to report legitimate work activities, the result could be an artifactually reduced rate of employment for that group.

For several vignettes, education and age are correlated with interpretations of work. Race and sex and (surprisingly) employment status are not. Table 3 shows that for the "accounting" and "antique shop" vignettes, education is positively associated with the correct "yes" answer. Responses to three vignettes ("antique shop," "painting for $20," and "hospital volunteer") are significantly related to age. For the "painting" vignette, the relationship is not monotonic. For the other two vignettes, younger respondents appear to hold a broader definition of work, and older ones have a narrower interpretation.

Table 3. Classification of work activities by age and education

Vignette			Completed years of education						
			1-8	9-11	12	13-15	16	17+	Total
(I)	Accounting (X^2 = 17.92, d.f.=5, p<.01)	% Yes	42	42	49	53	56	56	50
		N	130	236	804	413	222	172	1977
(K)	Antique (X^2 = 19.25, d.f.=5, p<.01)	% Yes	55	58	60	67	67	69	59
		N	129	234	791	408	217	170	1949

			Age in years							
			<24	25-29	30-39	40-49	50-59	60-69	70+	Total
(K)	Antique (X^2 = 32.91, d.f.=6, p<.01)	% Yes	60	69	64	62	56	55	46	59
		N	141	190	483	363	273	259	240	1949
(J)	Painting (X^2 = 16.83, d.f.=6, p<.01)	% Yes	63	57	64	70	66	68	58	64
		N	138	189	495	363	280	265	246	1976
(L)	Hospital (X^2 = 58.42, d.f.=6, p<.0001)	% Yes	59	51	40	35	33	29	32	38
		N	140	189	485	364	278	270	247	1973

In order to understand respondents' interpretations of work and evaluate the scale properties of the vignettes, we applied the Rasch measurement model to these data. We wished to address two questions:

> First, is there an underlying dimension of meaning which accounts for the diverse response patterns found in these data? Alternatively, do respondents apply different rules or criteria to classify each of these somewhat unusual situations, with no unifying concept governing their classifications? If this is true, interpreting these items as a scale of an underlying work concept is inappropriate.
>
> Second, to the extent that responses to these items are structured by a coherent concept, is the response structure constant across the population? In other words, do we find evidence of a shared concept of work that explains the pattern of associations among these five items?

The Rasch measurement model is well adapted to address such questions about the structure of responses to survey items. In this paper, we will focus primarily on the first question but comment on the second question in the concluding section. The particular methods applied here were developed for categorical data, and especially for dichotomous response data, by Goodman (1978), Duncan (1984), Tjur (1982), and others.

The Rasch measurement model postulates that several items are indirect measures of a single underlying attitude or trait that cannot be observed directly. (In this case, the latent variable is the propensity to classify situations as "work.") The Rasch model postulates that the odds of agreeing with each item is the product of an *item parameter*, and a *person parameter*. The item parameter in our application may be thought of as a measure of how closely the vignette matches the respondents' idea of "work." The person parameter reflects each individual's value on the underlying latent variable. It is assumed that the item parameters are constant across respondents, and that the person parameters do not depend on which question is being asked. This property is referred to as the separability of item and person parameters. (This discussion follows Duncan, 1984.)

In short, the model postulates an underlying trait, or dimension of meaning, that entirely accounts for the associations among responses to the particular vignettes. Conditional on the unobserved value of the underlying trait for a given respondent, the respondent's answers to the items are assumed to be statistically independent, with probabilities that are determined by the joint effect of the respondent's person parameter and by item parameters that do not vary over respondents.

In order to introduce the mathematical notation for the Rasch model, we first examine a three-way cross-classification of responses to the first three vignettes listed in Table 1, for which

the correct answer is "yes" in terms of the CPS definition––assisting in a family accounting business, being paid for painting, and setting up an antique store.

The first column of Table 4 presents the eight observed cell frequencies, excluding respondents with no answer on one or more items. The second shows the fitted values assuming statistical independence. This model can be written in a log-linear form as

$$\ln(F_{ijk}) = \lambda + \lambda^{I}_{i} + \lambda^{J}_{j} + \lambda^{K}_{k}, \quad (1)$$

where F_{ijk} represents the expected values under the log-linear model with the restrictions $\lambda^{I}_{2} = -\lambda^{I}_{1}$, $\lambda^{J}_{2} = -\lambda^{J}_{1}$, and $\lambda^{K}_{2} = -\lambda^{K}_{1}$.

Table 4. Observed and fitted frequencies for the cross-classification of work vignettes for accountant, painting, and setting up an antique business

Cells				(1)	(3)	(4)	(6)
Act.	Pnt.	Ant.	Observed	Indep.	All two-factor	Rasch	Rasch w/ accnt x antiq.
Y	Y	Y	538	362.19	541.26	538.00	538.00
N	Y	Y	253	349.91	249.74	273.84	249.20
Y	N	Y	185	200.34	181.74	132.10	185.00
N	N	Y	130	193.55	133.26	162.07	133.80
Y	Y	N	140	245.61	136.74	172.07	143.80
N	Y	N	264	237.28	267.26	211.10	264.00
Y	N	N	81	135.86	84.26	101.84	77.20
N	N	N	265	131.25	261.74	265.00	265.00
		X^2		341.08	0.49	52.61	0.45
		d.f.		4	1	2	1

This parametric form of the log-linear model may be found in Goodman (1978) and elsewhere. The λ parameters are analogous to coefficients of linear regression, where the object in this case is to predict the distribution of frequencies among the eight cells. The first λ parameter, analogous to an overall mean effect in linear regression, varies with the total sample. The second, λ^I_i, depends on level i of I, and similarly the other two parameters each correspond to a single variable.

Equation (1) expresses relationships between the expected values F_{ijk} of the cross-classification for the sample individuals. The model of statistical independence for F_{ijk} arises by assuming that responses to each of the three items are statistically independent and that the probabilities do not vary across sampled individuals. One way to express this, which is advantageous for comparison to the Rasch model, is to write

$$p_{ijk} = a^{(2-i)}b^{(2-j)}c^{(2-k)}/\{(1+a)(1+b)(1+c)\}, \qquad (2)$$

where p_{ijk} represents the probability of responses i, j, and k, for any individual in the sample. In fact, the parameters a, b, and c are related to (1) by $\lambda^I_1 = 0.5 \ln(a/(1+a))$, $\lambda^J_1 = 0 .5 \ln(b/(1+b))$, and $\lambda^K_1 = 0.5 \ln(c/(1+c))$.

Comparison of the observed and fitted values for equation (1) and the Pearson chi-square, X^2, of 341.09 on four degrees of freedom, both indicate a poor fit of this model. In particular, the observed counts in the two cells corresponding to consistent answers, 538 for all "yes" answers and 265 for all "no" answers, each greatly exceed the prediction under independence, 362.19 and 131.25, respectively.

The next model presented in the table includes all two-way interactions such that

$$\ln(F_{ijk}) = \lambda + \lambda^I_i + \lambda^J_j + \lambda^K_k + \lambda^{IJ}_{ij} + \lambda^{IK}_{ik} + \lambda^{JK}_{jk} \qquad (3)$$

with $\lambda^{IJ}_{11} = \lambda^{IJ}_{22} = -\lambda^{IJ}_{12} = -\lambda^{IJ}_{21}$, etc. Model (3) fits the data quite well, with $X^2 = 0.49$ on one degree of freedom. In this model, interaction is allowed between each pair of variables. The single residual degree of freedom is due to the omission of the three-way interaction term λ^{IJK}_{ijk} from the model; otherwise, (3) would be "saturated" and fit the observed data exactly.

The next model, labelled "Rasch" in the table, represents the fit of a log-linear model of the form

$$\ln(F_{ijk}) = \lambda + \lambda^{I}_{i} + \lambda^{J}_{j} + \lambda^{K}_{k} + \lambda^{R}_{r} \qquad (4)$$

where r = 0, 1, 2, or 3, represents the number of "yes"answers to items I, J, and K. In addition to the usual restrictions on the λ values, the additional restrictions

$$\lambda^{I}_{1} + \lambda^{J}_{1} + \lambda^{K}_{1} = 0 \qquad (5)$$

and

$$\sum_{r} \lambda^{R}_{r} = 0$$

are imposed in order to identify the parameters. Model (4) arises as the log-linear model asymptotically equivalent to the estimation of the item parameters of the Rasch (1960/1980) model, conditional on the distribution of subject parameters. Model (4) and generalizations to an arbitrary number of variables were established by Tjur (1982), who showed that the log-linear model (4) for the expected F_{ijk} arises when the probability distribution for each individual takes the form initially characterized by Rasch, i.e.

$$p_{ijk} = (ax)^{2-i}(bx)^{2-j}(cx)^{2-k}/\{(1+ax)(1+bx)(1+cx)\}$$

where x represents a parameter for each individual and a, b, and c still represent parameters for variables I, J, and K. This equation closely parallels (2). If x was the same for each member of the population, then the expected values could still be expressed in the same form as equation (1). When x varies among individuals, however, the effect is to jointly increase or decrease

individuals' probabilities of "yes" responses for I, J, and K. Variation in an unobserved x thus may account for specific patterns of dependence among the observed I, J, and K. The presence of the term λ^R_r in equation (4) is to represent variation in x among individuals in the population (Duncan, 1984).

The Rasch parameter expresses heterogeneity among individuals in the latent propensity to view different situations as "work." It is the assumption that once variation among individuals in the underlying trait is explicitly accounted for by including the Rasch parameter in the model, then associations among individual response variables should vanish. The fit of the Rasch model is considerably better than that for the independence model and, because of the form of equation (4), the cells with all "yes" answers or all "no" answers are fitted exactly under the model. Nonetheless, model (4) fits poorly, with $X^2 = 52.61$ on 2 degrees of freedom.

A compromise between models (3) and (4) is possible, taking the form

$$\ln(F_{ijk}) = \lambda + \lambda^I_i + \lambda^J_j + \lambda^K_k + \lambda^{IK}_{ik} + \lambda^R_r \qquad (6)$$

This model is similar to model (4) but includes one of the three interaction terms of model (3), "accounting" by "antique." The notion of including such interactions was mentioned briefly by Duncan in Turner and Martin (1984: p. 226), but the theory and implications of such additions have been provided in detail by Fay and Turner (1989). Model (6) provides quite a good fit, with $X^2 = 0.45$ on one degree of freedom. It postulates that the tendency for subjects to respond "yes" to each of the items is influenced by a single underlying propensity; however, an additional term is included in the model to account for an association between the corresponding two variables, above and beyond the underlying propensity to say "yes" or "no" affecting all three.

In place of the interaction term between the I and K variables in model (6), it is possible instead to substitute an interaction term involving either of the other two possible pairs of

variables. The chi-square tests for these other two possibilities are extremely less favorable, however, being 30.28 and 41.42, each on a single degree of freedom. Thus, the specific choice of interaction term in model (6) is highly critical in this application.

Table 5 presents the estimated parameters for models (3), (4), and (6), although without displaying parameters for the λ^R_r. The three parameters corresponding to two-way interactions in model (3) are all positive and statistically significant, with the interaction between "accounting" and "antique" substantially more pronounced than the other two interaction terms.

Table 5. λ Parameters for three models for work vignettes for accountant, painting, and setting up an antique business

	(3) All two-factor		(4) Rasch		(6) Rasch w/ accnt x antiq.	
	λ	(s.e.)	λ	(s.e.)	λ	(s.e.)
Accnt.	−0.090	(.026)	−0.199	(.022)	−0.219	(.023)
Paint.	0.278	(.020)	0.166	(.025)	0.163	(.022)
Antiq.	0.175	(.027)	0.033	(.016)	0.056	(.019)
Accnt. x Paint.	0.116	(.023)				
Accnt. x Antiq.	0.361	(.030)			0.233	(.031)
Paint. x Antiq.	0.152	(.025)				

Although model (4) does not fit the data well, Table 5 presents the item parameters of this model. As noted earlier, the model is expressed with the constraint of equation (5) that the sum of the item parameters be forced to zero. The ordering of item parameters indicates that the painting vignette elicits the most "yes" responses and accounting the least. The parameters of model (6) can be seen to be closely related to the parameters of model (3) in this application but

are affected by the constraint of equation (5).[2]

The positive value for λ^{IK}_{11} in model (6), estimated as 0.233, represents an additional level of association between I and K, beyond that suggested by the Rasch model in model (4). Unpaid contributions to a family business and setting up an antique shop may seem especially similar to respondents because both activities pertain to business activities without direct and immediate cash benefit to the person in the vignette. The particularly high association between these vignettes may arise if many of the respondents use direct cash payment as a rule or heuristic for judging whether a person is working.

4 Expanding the Model

We turn next to the question of whether the remaining two vignettes–volunteering in a hospital and giving blood–for which the correct answer is "no," can be incorporated into a model similar to model (6). To do so, we first analyze the five-way cross-classification of all five vignettes, again setting aside any missing responses, with n=1771.

We begin by considering two different scorings for the five vignettes: one with all items scored by "yes"/"no" and one scored by the correctness of the respondent's answer. (In the second scoring, "no" is correct for the hospital volunteer and blood donor, and "yes" is correct for

---

[2] The main effect or item parameters of model (4) can be obtained almost exactly by subtracting 0.120 from the main effect parameters of model (2) (i.e. −0.090 −0.120 = −0.210 vs. −0.219; 0.278 − 0.120 = 0.158 vs. 0.163; 0.175 − 0.120 = 0.055 vs. 0.056). Similarly, two of the interaction terms of equation (2) disappear from model (4), but the remaining term of equation (2) reappears in model (4) changed by an amount almost exactly equal to the average of the difference of the two other terms (i.e. 0.361 − (0.116 + 0.152)/2 = 0.227 vs. 0.233). None of these observations is coincidental but instead represents relationships that should be expected to hold quite closely, although generally not exactly, in contexts in which both equations (2) and (4) fit the observed data quite well (Fay and Turner, 1989).

the other three.) Table 6 presents the parameters for factorial log-linear models with all possible two-way interactions for the two different scorings. The effect of the "correctness" scoring is to reverse signs for specific parameters compared to the "yes" scoring. Two-way associations involving hospital volunteer with each of the first three variables show positive associations under the "yes" ordering, but negative associations under the ordering by correctness.

Table 6. λ parameters for the model including all two-factor interactions, ordered by "yes" and "correctness"

	"Yes" ordering		"Correctness" ordering	
	λ	(s.e.)	λ	(s.e.)
Accnt	0.020	(.037)	0.020	(.037)
Paint.	0.630	(.056)	0.630	(.056)
Antiq.	0.226	(.033)	0.226	(.033)
Hosp.	−0.300	(.036)	0.300	(.036)
Blood	−1.131	(.065)	1.131	(.065)
Accnt x Paint	0.096	(.023)	0.096	(.023)
Accnt x Antiq.	0.311	(.031)	0.311	(.031)
Accnt x Hosp.	0.411	(.035)	−0.411	(.035)
Accnt x Blood	−0.022	(.036)	0.022	(.036)
Paint x Antiq.	0.144	(.024)	0.144	(.024)
Paint x Hosp.	0.062	(.034)	−0.062	(.034)
Paint x Blood	0.430	(.052)	−0.430	(.052)
Antiq. x Hosp.	0.148	(.025)	−0.148	(.025)
Antiq. x Blood	0.023	(.031)	−0.023	(.031)
Hosp. x Blood	0.061	(.042)	0.061	(.042)

Responses for the blood donation vignette show a less consistent pattern. Again, the ordering by "yes" responses produces the most consistent results but the only appreciable association with any other variable is a strong association with painting. Notably, it is in these two vignettes that money explicitly changes hands, so the association may be the effect of using the exchange of money as a heuristic to answer both questions.

The results in Table 6 imply there is a general propensity for respondents consistently to prefer or avoid "yes" responses, rather than a propensity to give correct or incorrect answers. Thus, if either or both of the remaining variables are to be incorporated into the Rasch model, the only suitable ordering is one in which all variables are ordered by "yes"/"no."

Since the results in Table 6 imply a generally higher association for responses to hospital volunteer with the first three variables than for responses to blood donor, the hospital volunteer item would appear a more suitable candidate for inclusion in the Rasch model. Several observations support this intuition and there is a relatively simple model which fits a four-way table of the first three variables crossed by hospital volunteer:

$$\ln(F_{ijk\ell}) = \lambda + \lambda^{I}_{i} + \lambda^{J}_{j} + \lambda^{K}_{k} + \lambda^{L}_{\ell} + \lambda^{IK}_{ik} + \lambda^{IL}_{i\ell} + \lambda^{JL}_{j\ell} + \lambda^{R}_{r}, \qquad (7)$$

This is obtained by modifying model (6) by the inclusion of two additional pairwise interaction terms, with $X^2 = 5.76$ on five degrees of freedom. One interaction $\lambda^{JL}_{j\ell}$ is slightly negative, i.e. -0.088. An analogous four-way table of the first three variables by blood donor is fitted by similar model in which all three interaction terms between blood donor and each of the other variables are included, with $X^2 = 1.90$ on 4 degrees of freedom; each interaction makes a significant contribution to the model, but two have estimated values of -0.163 and -0.115. The two negative interactions are evidence that the blood donor variable, except for its association with the painting variable, does not share the same dimension underlying the first three.

Similarly, if all five variables are included as response variables in a single model by adding to equation (7) interaction terms between blood donor and each of the other four variables as well as a λ^{M}_{m} item parameter, then three of the four new interactions assume negative values, i.e. −0.131, −0.197, and −0.122. Although the fit of this model is quite acceptable, with X^2 = 10.47 on 15 degrees of freedom, the negative outcomes for interactions with three of the four other variables, each of which is statistically significant, again imply that the blood donor vignette does not tap the same underlying dimension. Therefore, we conclude that model (7), for four response variables excluding blood donor, best describes the data.

As shown in Table 7, model (7) adds two interactions to model (6). The positive interaction between the accountant and hospital volunteer vignettes may indicate that many respondents classify them similarly because no cash is expected from either activity, including any time in the future. The negative association between painting and hospital volunteer is weak and may occur because the painter receives the most immediate cash reward and the hospital volunteer receives none.

Table 7. Parameters of the best-fitting model (7) describing responses to four work vignettes

	λ	(s.e.)
Accnt	−0.010	(.029)
Paint.	0.242	(.022)
Antiq.	0.157	(.019)
Hosp.	−0.389	(.021)
Accnt x Antiq.	0.201	(.034)
Accnt x Hosp.	0.274	(.042)
Paint. x Hosp.	−0.088	(.041)

The contribution of the λ^R_r parameters to model (7) is illustrated in Table 8, which compares the observed frequencies with the expected frequencies of 0, 1, 2, 3, or 4 "yes" answers under the log-linear model in the form of model (7) without the Rasch parameters and with the expected frequencies under the independence model. The corresponding expected frequencies of "yes" answers under model (7) and the Rasch model without interactions agree with the observed frequencies. Without the λ^R_r parameters, the log-linear model, and especially the independence model, predict far too few responses for all "yes" or all "no." The role of the λ^R_r parameters is to represent the effect, through the Rasch model, of heterogeneity in the population in the propensity to classify situations as "work."

Table 8. Comparison of observed and fitted values for model (7) with and without λ^R_r

Number of "yes" responses	Observed and fitted with λ^R_r	Fitted without λ^R_r	Independence
4	352	294.98	133.76
3	365	424.10	510.99
2	470	472.21	688.40
1	393	439.30	391.15
0	224	173.38	79.69

5 Conclusion

Our data are consistent with the assumption that there is an underlying dimension of meaning which accounts for responses to four of five of the work vignettes. Respondents appear to vary along a dimension of inclusiveness. At one extreme are respondents who hold a strict definition, and at the other extreme are respondents who appear willing to include marginal activities as work. Only four of five items can be considered as measures of the underlying latent

variable. Giving blood for money does not "fit" with the others as a scale.

Vignettes such as those designed for this study can be useful in two phases of questionnaire design. First, they are useful as diagnostic tools, to identify problems of respondent comprehension of existing questions. By using vignettes, one can learn how a survey term or concept is applied in practice to situations of interest to the designer. The methodology preserves contextual validity because respondents are asked to interpret the term or concept in the context of CPS reporting, and not in general or in a different context. Information about which words and phrases are misunderstood and which types of situations are misclassified can be used by the questionnaire designer to reword questions, add instructions or examples, etc., or, alternatively, to revise the survey definitions to bring them more in line with respondents' interpretations. In this application, the examples given in the vignettes highlight some of the problems we have attempted to correct by redesigning the CPS questionnaire to create greater uniformity in respondents' reporting of work activities, and increase respondents' conformity with the technical definition used in the survey (see Copeland and Rothgeb (1990) for a description of questionnaire revisions and testing plans).

In the second phase of questionnaire redesign, vignettes are useful to test and compare respondent comprehension of one or more experimental questionnaires with a control questionnaire. Assuming other factors are constant, the preferred questionnaire usually would be the one which is best understood by respondents, as measured by their performance on the vignette scale as well as other measures and criteria for improvement.

Rasch analysis of the scale properties of the five vignettes sheds new light on the meaning of "work." It is significant that this ordinary and seemingly simple construct has diverse interpretations in the population. It is also instructive that the dimension underlying responses to

the vignettes is *inclusiveness* and not *correctness*. This may have implications for how questionnaire designers go about their work, because it may mean that efforts to increase respondent accuracy are naive. It may be that surveys can be designed to induce respondents to be more inclusive or more restrictive in their reporting, but not more accurate, in terms of a precise survey definition. Perhaps success at reducing either type of error (over-reporting or under-reporting) invariably increases the other, so that the price the survey designer pays for more complete reporting is over-reporting. If under-reporting and over-reporting errors trade off in this fashion, then one might opt for a design strategy which casts a "broad net" (cf. Biderman, 1980) and elicits full reporting based on an inclusive definition, building in questions and probes to screen out invalid reports. Typically it is easier to identify and eliminate invalid reports than to correct for omissions.

The dominance of the inclusiveness dimension also has implications for the design of similar vignette studies. If the interest in is both inclusiveness and correctness, then it is advantageous to design a set of vignettes such that half are “yes” items. Without a balance between “yes” and “no” items, the study could confound correctness with inclusiveness.

As noted earlier, we were interested both in whether the vignettes could be summarized by a single dimension and whether the patterns of association were consistent across the population. Using an expanded form of the model, we were able to detect some differences in patterns by age and education, for example, such that college-educated respondents exhibit more intense interactions between some pairs of items. We also found that a model similar to model (7) described patterns of association for CPS interviewers, particularly those with less than five years’ experience, when they were asked the vignettes.

We do not yet know much about the error properties of the vignettes, which may

themselves be subject to various unintended interpretations. In particular, respondents may classify the work vignettes erroneously because the situations are too hypothetical and remote from their own experience. We examined this question for two vignettes for which we could identify respondents to whom the situation described in the vignette was applicable. For one vignette, respondents for whom the situation was relevant were more likely to be correct than those for whom it was not relevant, and for the other vignette, the reverse was true (see Campanelli, Martin, and Creighton, 1989).

Our results suggest that the Rasch measurement model is useful for designing and refining scales, such as our five-item scale measuring the meaning of work. The ideal of scale construction is to design a set of items which tap a single dimension, and the Rasch model provides a rigorous test of the scalability of items. In our case, fitting the data to a Rasch model revealed that the blood donation vignette does not fit in the scale and we dropped it. We also included interaction terms between pairs of items to account for associations stronger than the original Rasch model would predict. High associations of this sort are also undesirable in scale construction, but analyzing a Rasch model with interactions is useful in vignette analysis to uncover patterns of association beyond a single, underlying dimension, as a means to discover whether respondents use some specific strategies for certain classes of problems.

Acknowledgments

This paper reports the results of research undertaken by Census Bureau staff. The views expressed are attributed to the authors and do not necessarily reflect those of the Bureau of the Census. The authors with to thank O. D. Duncan for helpful comments on an earlier version of this paper.

References

Bailar, B. A., & Rothwell, N. D. (1984) Measuring employment and unemployment, in: C. F. Turner and E. Martin (Eds), *Surveying Subjective Phenomena,* Vol. 2, pp. 129-142 (New York, Russell Sage).

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

Biderman, A. D. (1980) Crime-circumscribed vs. broader-net screening approaches, *Item 361 in Crime Survey Research Consortium Teleconference* (Washington, DC, Bureau of Social Science Research).

Biderman, A. D., Cantor D., Lynch, J. P., & Martin, E. (1986) *Final Report of the National Crime Survey Redesign Program* (Washington, DC, Bureau of Social Science Research).

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

Campanelli, P. C., Martin, E. A., & Creighton, K. P. (1989) Respondents' understanding of labor force concepts: Insights from debriefing studies, in: *Proceedings of the Fifth Annual Research Conference*, pp. 361-374 (Washington, DC, Bureau of the Census).

Copeland, K. & Rothgeb, J. (1990) Testing alternative questionnaires for the Current Population Survey, *Proceedings of the Survey Research Methods Section* (Alexandria VA, American Statistical Association).

Duncan, O. D. (1984) Rasch measurement in survey research: Further examples and discussion, in: C. F. Turner and E. Martin (Eds), *Surveying Subjective Phenomena,* Vol. 2, pp. 367-403 (New York, Russell Sage Foundation).

Fay, R. E. & Turner, C. F. (1989) *RASCHPLX: Rasch analysis for contingency tables with complex sample designs*, unpublished program documentation, (Washington, DC, National Academy of Sciences).

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

Martin, E. A. (1987) Some conceptual problems in the Current Population Survey, in: *Proceedings of the Survey Research Methods Section*, pp. 420-424 (Alexandria, VA, American Statistical Association).

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

Mishler, E. G. (1986) *Research Interviewing: Context and Narrative* (Cambridge, MA, Harvard University Press).

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

Rossi, P. H. & Nock, S. L. (Eds) (1983) *Measuring Social Judgements: The Factorial Survey Approach* (Beverly Hills, CA, Sage).

Rothgeb, J. M. (1982) Summary report of July follow-up of the unemployed, memorandum for K.A. Riccini, December 20, 1982, (Washington, DC, U.S. Bureau of the Census).

Suchman, L. & Jordan, B. (1990) Interactional troubles in face-to-face survey interviews, *Journal of the American Statistical Association* 25, pp. 232-253.

Schwarz, N. (1987) Cognitive aspects of labor surveys in a multi-national context, paper presented at the *Meeting of the Working Party on Labor Statistics, OECD, Paris*.

Stepick, A., & Stepick, C. (1990) What's in it for me? What's in it for you?, *Ethnographic*

Research Report 11, (final report for Joint Statistical Agreement 88-26 between the Bureau of the Census and Florida International University).

Tjur, T. (1982) A connection between Rasch's item analysis model and a multiplicative Poisson model, *Scandinavian Journal of Statistics*, 9, pp. 23-30.

Turner, C. F. & Martin, E. (Eds) (1984) *Surveying Subjective Phenomena,* Vol. 1 (New York: Russell Sage.

Wolfgang, M. E., Figlio, R. M., Tracy, P. E., & Singer, S. I. (1985) *The National Survey of Crime Severity, NCJ-96017* (U. S. Department of Justice, Bureau of Justice Statistics, Washington, DC, U.S. Government Printing Office).