October 2005

VIGNETTES AND SELF-REPORTS OF WORK DISABILITY IN THE US AND THE NETHERLANDS*

JEL codes: J28, I12, C81

Keywords: Work limiting disability, Vignettes, Reporting bias

High and rising rates of work disability are a pervasive problem in many industrialized countries (See Robert Haveman and Barbara Wolfe, 2000, or John Bound and Richard Burkhauser, 1999). The fraction of workers reporting work disability is vastly different across countries with similar levels of economic development and comparable access to modern medical technology and treatment. Institutional differences in eligibility rules or generosity of benefits no doubt explain part of the differences in disability rolls (see, Bound and Burkhauser, 1999, Burkhauser and Mary Daly, 2002, and Thomas DeLeire, 2000). However, recent survey data show that significant differences between countries are also found in self-reports of work limiting disabilities. In comparing such self-reports, a basic question concerns the extent to which people living in the same or in different countries use the same response scales when they answer questions about work disability. If they use the same scales, differences in reported work disability reflect true differences across countries in disabilities affecting work. But if response scales differ systematically, adjustments must be made before conclusions about international differences in true work disability can be drawn.

The paper puts forth a new approach to the measurement of work disability. We utilize a vignette methodology to evaluate how people within and across countries set thresholds that result in labeling some people work disabled while other people are not so described. Our vignette questions ask respondents to evaluate on the same scale on which they evaluate themselves the severity of work disability problems of hypothetical scenarios and people.[1]

While we use vignettes for work disability, our methodology applies to many economic applications with subjective scales. Gary King's web-site (http://gking.harvard.edu/vign/eg/) lists vignettes pertaining to a variety of domains, including health, health care, school community strength, HIV risk, state effectiveness, and corruption. Within economics one could use vignettes

to address differences in perceptions of risk, opinions about income inequality and poverty, workload and pay, valuation of the quality of public services, etc.

This research performs an international comparison of two countries: US and The Netherlands. These countries differ in several relevant dimensions—observed rates of self-reported work disability, the generosity of government programs providing income support for people with a work disability, and perhaps national norms about the appropriateness of not working when work disabled (see, Leo Aarts et al. 1996). But one might reasonably suspect that these two countries differ less in the 'objectively' measured health status of the population. For this reason, we believe that this international comparison is particularly useful in understanding the salient research issues that have dominated the scientific literature on work disability.

A unique aspect of the research is that we address these issues in a classic random experimental form. This is because we have access to Internet samples in both countries allowing us to place experimental disability vignettes modules into these panels. These samples are the Dutch CentERpanel and the RAND MS Internet panel for the US, both described in detail below.

This paper is organized as follows. In the next section, we describe the vignette methodology and our statistical model that corrects for response scale differences across countries. Section II briefly discusses our data and some measurement issues, and presents descriptive statistics on self-reported work limitations in the US and The Netherlands. Section III presents the empirical results and their implications for interpreting observed differences in work disabilities. Conclusions follow in Section IV.

I. The Vignette Methodology - *A. The Intuition about Vignettes*

In this section, we provide an intuitive description of the use of vignettes for identifying response scale differences and then sketch our statistical approach. The basic idea is illustrated in

Figure 1, which presents the distribution of health in two hypothetical countries. The density of the continuous health variable in country A is to the left of that in country B, implying that on average, people in country A are less healthy than in country B. The people in the two countries, however, use very different response scales if asked to report their health on a five-point scale (poor-fair-good-very good-excellent). In this figure, people in country A have a much more positive view of a given health status than people in country B. Someone in country A with health indicated by the dashed line would report to be in very good health, while a person in country B with the same actual health would report "fair." The frequency distribution of self-reports in the two countries would suggest that people in country A are healthier than those in country B—the opposite of the true health distribution. Correcting for differences in the response scales (DIF, "differential item functioning," in the terminology of Gary King et al., 2004) is essential to compare the actual health distributions in the two countries.

Vignettes can be used to do the correction. A vignette question describes the health of a hypothetical person and then asks the respondent to evaluate that person's health on the same scale used for a self-report on their own health. Since vignette descriptions are the same, vignette persons in the two countries have the same actual health. For example, respondents can be asked to evaluate the health of a person whose health is given by the dashed line. In country A, this will be evaluated as "very good." In country B, the evaluation would be "fair." Since actual health is the same in the two countries, differences in country evaluations must be due to DIF.

Vignette evaluations thus help to identify the differences between the response scales. Using scales in one country as the benchmark, the distribution of evaluations in the other country can be adjusted by evaluating on the benchmark scale. The corrected distribution can be compared to that in the benchmark country—both are on the same scale. In the example in the figure, this

leads to the correct conclusion that people in country B are healthier than those in country A, on average. The underlying assumption is *response consistency:* each respondent uses the same scale for the self-report and the vignette evaluations. King et al. (2004) provide evidence supporting this assumption by comparing self-reports of vision corrected and not corrected for DIF using vignette evaluations with an objective measure of vision.

We will apply the vignette approach to work limiting disability, using vignettes not only to obtain international comparisons corrected for DIF, but also for comparisons of different groups within a given country. For example, it is often hypothesized that men self-report themselves in better health than objective circumstances would warrant, that as they age people adjust their norms about what constitutes good health downward, and that some of the SES health gradient reflects different health thresholds by SES rather than true health differences. Vignettes offer the potential for systematic testing of these hypotheses.

B. Formal Model with Vignettes on Work Limiting Disability

Our model explains respondents' self-reports on work limitations and their reports on work limitations of hypothetical vignette persons. The first of these is the answer, on a five-point scale (Y_{ri} for respondent i), to the question: *"Do you have any impairment or health problem that limits the kind or amount of paid work you can do?"* Self-reports are modeled as a function of respondent characteristics X_i (including a country dummy and interactions with that dummy) and an error term ε_{ri} by the following ordered response equation:

$$Y^*_{ri} = X_i\beta + \varepsilon_{ri};\ \ \varepsilon_{ri} \sim N(0,\sigma^2_r),\ \varepsilon_{ri} \text{ independent of } X_i \tag{1.1}$$

$$Y_{ri} = j \text{ if } \tau_i^{j-1} < Y^*_{ri} \le \tau_i^j,\ \ j = 1,...5 \tag{1.2}$$

The thresholds τ_j^i between the categories are given by

$$\tau_i^0 = -\infty,\ \tau_i^5 = \infty,\ \tau_i^1 = \gamma^1 X_i + u_i,\ \tau_i^j = \tau_i^{j-1} + \exp(\gamma^j X_i),\ j = 2,3,4$$
$$u_i \sim N(0, \sigma_u^2),\ u_i \text{ independent of } X_i \text{ and the other error terms in the model} \quad (1.3)$$

The fact that different respondents use different response scales τ_i^j is called "differential item functioning" (DIF). The term u_i introduces an unobserved individual effect in the response scale. It implies that evaluations of different vignettes are correlated with each other and with the self-reports (conditional on X_i), since some respondents will tend to use high thresholds and others will use low thresholds in all their evaluations.

Define a benchmark respondent with characteristics $X_i = X(B)$. The DIF adjustment involves comparing Y_{ri}^* to thresholds τ_B^j rather than τ_i^j, where τ_B^j is obtained in the same way as τ_i^j but using $X(B)$ instead of X_i. A respondent's work ability is computed using a benchmark scale instead of a respondent's own scale. This does not give an adjusted score for each individual (since Y_{ri}^* is not observed) but it can be used to simulate adjusted *distributions* of Y_{ri} for the whole population or conditional upon some of the characteristics in X_i.

Using self-reports on own work disabilities only, parameters β and γ^1 are not separately identified,[2] only their difference. For example, consider country dummies: people in two different countries can have systematically different work disability, but if the scales on which they report their work disability can also differ across countries, then self-reports are not enough to identify the work disability difference between the countries.

In the surveys we use, each respondent answered $L=15$ vignette questions, five in each of the three domains affect, pain, and heart problems. The evaluations Y_{li} of vignettes $l=1,...,L$ are modeled using similar ordered response equations:

$$Y_{li}^* = \theta_l + \theta^d \text{Female}_{li} + \varepsilon_{li} \quad (1.4)$$

(1.5) $Y_{li} = j$ if $\tau_i^{j-1} < Y_{li}^* \leq \tau_i^j$, $j = 1,...5$

(1.6) $\varepsilon_{li} \sim N(0, \sigma^2)$, independent of each other, of ε_{ri} and of X_i

Apart from dummies to indicate the vignettes, the only explanatory variable in (1.4) is a dummy for the gender of the vignette person. It is included because preliminary analysis suggested that respondents react differently to vignettes with a female name than with a male name.[3] We allow these differences vary by domain, i.e., we include separate coefficients θ^d for *d=affect (l=1,...,5), pain (l=6,...,10)* or *CVD (l=11,...,15).* The assumption of *"response consistency"* means that the thresholds τ_i^j are the same for the self-reports and the vignettes.

With these assumptions, it is clear how vignette evaluations can separately identify β and γ $(=\gamma^1,...,\gamma^5)$: From the vignette evaluations alone, γ, θ, $\theta_1,...\theta_5$ can be identified (up to the usual normalization of scale and location). From the self-reports, β can then be identified in addition. Thus the vignettes can be used to solve the identification problem due to DIF.

II. Data and vignette evaluations - *A. Data Sources*

We use information obtained from two Internet surveys, which we conducted in both countries. For The Netherlands, we use the CentERpanel, which includes about 2,250 households who have agreed to respond to questions every weekend over the Internet. This Dutch sample is not restricted to households with their own Internet access. Respondents are recruited by telephone. If they agree to participate and do not have Internet access, they are provided with Internet access (and if necessary, a set-top box). Thus, CentERpanel is representative of the adult Dutch population except the institutionalised.[4] The sample we use for estimation has 1,977 respondents who participated in several interviews with questions on work disability in 2003.

From multiple waves collected in the past, CentERpanel has a rich set of variables on demographic and labor market characteristics of respondents, as well as several salient

dimensions of health. In August 2003, we collected work disability self-reports and vignette evaluations (described below). The Internet infrastructure makes CentERpanel an extremely valuable tool to conduct experiments, with possibilities for randomization of content. Production lags are very short, with about one month between module design and data delivery. Based on our initial analysis, we fielded a second wave in October allowing us to randomize the gender used in vignette questions. (In the first wave, vignette persons always had the same gender as the respondent.) A third wave of experiments was administered in December 2003.

The RAND MS Internet panel was recruited from respondents age 40 plus in the Monthly Survey (MS) of Michigan's Survey Research Center. The MS, the leading consumer sentiments survey, produces the widely used Index of Consumer Attitudes. MS respondents age 40 plus are asked if they have Internet access and, if yes, if they are willing to participate in Internet surveys. Those who agree are added to our household panel to be interviewed regularly over the Internet. Our sample consists of 672 respondents interviewed in the first six months of 2004. Because of the smaller size of the RAND MS Internet sample, we also use 15,740 respondents younger than 75 in the 1998 wave of the Health and Retirement Study (HRS), the most recent wave with a representative cohort interviewed at ages 51-61. The HRS sample has self-reports on work disability like the RAND MS Internet panel and the CentERpanel, but has no vignette questions.

The question on work disability in the US and Dutch Internet surveys is: *"Do you have any impairment or health problem that limits the kind or amount of paid work you can do?"*

This question also appears in PSID and HRS. US respondents answer on a two-point scale (*yes* or *no*) while the possible answers are arrayed on the following five-point scale in the first wave of the Dutch survey: (1) no, not at all, (2) yes, I am mildly limited, (3) yes, I am moderately limited, (4) yes, I am severely limited, (5) yes, I am extremely limited, cannot work.

A few months later, Dutch respondents got the same work disability question but now with the two-point scale (*yes* or *no*). The top panel of Table 1 shows reported US disability rates by age from the PSID and Dutch disability rates obtained from CentERpanel using this two-point scale. Especially for middle age workers self-reported work disability is much higher in the Netherlands than in the US. This difference is almost twenty percentage points among the 45-54 years old. In contrast to this, the remaining rows in Table 1 suggest that the Dutch population is healthier than the US population. With the sole exception of the somewhat subjective domain of emotional problems, among those in the critical age groups of 45-64 years old, disease prevalence is always higher in the US.[5] It is not central to our argument that the Dutch are healthier than the Americans; the main point is that differences in standard measures of health are unlikely to explain the much higher Dutch self reported work disability rates.

B. Descriptive Statistics on Vignettes

We gave the Dutch and US Internet respondents five vignettes each in three domains of work disability—pain, affect (or emotional problems), and heart disease. These domains were chosen because preliminary analysis indicated that they were the most important determinants of work disability and because they span the range from a very subjective health condition (affect) to a quite objective one (heart disease) with pain being the intermediate case. Our hypothesis was that variation in response scales would increase with the subjectivity of the health problem.

Two examples, the least and most severe from the pain domain, are given below:[6]

1. [Katie] occasionally feels back pain at work, but this has not happened for the last several months now. If she feels back pain, it typically lasts only for a few days.
2. [Mark] has pain in his back and legs, and the pain is present almost all the time. It gets worse while he is working. Although medication helps, he feels uncomfortable when moving around , holding and lifting things at work.

All vignettes were presented with either a female or male name, randomized across respondents. Within each domain vignettes were presented in random order to eliminate any order effects. Comparing the rank ordering of vignette evaluations across respondents shows that different respondents tend to order vignettes in the same way. We designed the vignettes with an eye on cross-cultural comparability. Conceivably, one or two of the vignettes will still not be perfectly comparable, but that would only have a minor effect on the analysis, which will use 15 vignettes.

Table 2 compares Dutch and US vignette evaluations. In each domain the ordering of severity of the vignette description goes from least (=1) to most severe (=5).[7] Although health conditions of persons in the vignettes are the same in both countries, there are some substantial differences in the evaluation frequencies. In particular for the first two pain and affect vignettes describing people with relatively mild work limitations, US respondents much more often report that these persons have no limitation at all, where Dutch respondents have a larger tendency to use the intermediate categories "mildly" and "moderately." As hypothesized, the differences between the two countries in vignette evaluations are smallest in the more 'objective' domain of heart disease. There is some indication that the Dutch are reluctant to use the most extreme labeling of work disability (extremely) and are more comfortable in the middle. US respondents often evaluate a person as severely or extremely limited, where the Dutch still tend to answer "moderately." These patterns imply that the Dutch seem harder on vignette persons with a serious limitation and softer on those with a minor limitation. This tendency for the Dutch to run to the center is not limited to work disability. In a five point scale describing their general health status from excellent to poor, compared to the Dutch, US respondents are four times more likely to state that they are in excellent health, and twice as likely to say they are in poor health.

For a dichotomous question on whether one is work disabled or not, being softer on those with a minor condition is much more important than being harder on those with a serious work limitation. Whether one labels someone as 'severely' or 'extremely' work limited does not matter on a two-point scale, as people in both categories will always be seen as having a work disability. In contrast, the general reluctance of the Dutch (relative to the Americans) to say that someone is 'not at all' work limited is critical. Thus Table 2 suggests that especially in the domains of pain and affect the Dutch would be harder on themselves if they used the US scales. Using the US scales would reduce Dutch self-reported work disability prevalence, thereby reducing the difference in work disability prevalence between the two countries.

III. Empirical Estimates - *A. Model Specifications*

To estimate the model comparing work disability in the US and The Netherlands, three data sets are combined: the Dutch CentERpanel (waves 1, 2 and 3, in August, October and December 2003), the US RAND MS Internet panel, and the US HRS wave 1998.[8] CentERpanel and RAND MS have exactly the same vignette questions on pain problems, emotional problems, and cardio-vascular disease. HRS has no vignettes.

CentERpanel has self-reports on work disability on a five-point scale and on a two-point scale, while both US surveys have the two-point scale only. To link the US (and NL) self-reports on the two-point scale to the US (and NL) vignette evaluations on a five point scale, we expand the model described above with a transformation from the five-point scale to the two-point scale. The Dutch data with both scales for the same respondents show that the cut-off point between "yes" and "no" for the two-point scale is between the cut-off points between "no" and "mildly" and "mildly" and "moderately" for the five-point scale. Thus, we model the cut-off point $\tau_i(2)$ on the two-point scale as a weighted mean of the two first cut-off points on the five-point scale:[9]

(3.1) $$\tau_i(2) = \lambda \tau_i^1 + (1-\lambda)\tau_i^2$$

We assume the weight λ does not vary with individual characteristics and is the same in both countries. Thus the thresholds on the five-point scale and the thresholds on the two-point scale can have completely different structures in the two countries, but the relation between them is the same. The parameter λ is identified from the Dutch self-reports on both scales and applied to the US respondents. All parameters are estimated simultaneously by Maximum Likelihood

The equations for work disability and for the thresholds include a complete set of interactions with the country dummy for The Netherlands. Vignette equations and the scale transformation discussed above have no respondent characteristics, not even a country dummy.

B. Parameter Estimates and Within Country Implications

Within this basic structure we consider several models to test the sensitivity of our main results to different assumptions. The first "benchmark model" uses all 15 vignettes covering three domains (affect, pain and cvd), assuming a common response scale across these domains. [10]

Table 3 presents results for the work disability equation in this model, comparing it with a model not allowing for any threshold variation across respondents. The work disability equation in the latter model (first two columns) is almost identical to a standard probit for the probability of reporting a work disability, not taking account of potential differences in reporting thresholds. The remaining columns concern the model allowing for different response scales. The middle two columns are the estimated effects of respondent characteristics on the first response threshold γ^1, the most important threshold for determining whether someone claims to be work disabled on a two point scale (see (1.3) and (3.1)). The final two columns are the coefficients in the work disability equation after correcting for differential response scales. A likelihood ratio test strongly rejects the model not allowing for response scale variation against

the more general model that does allow for DIF. The same is found for each country separately, in line with the many significant parameters in the first threshold equation in Table 3.

We first focus on the impact of differential response scales on within country variation in work disability. In the model without the DIF correction, work disability in the US falls significantly with education level, rises with age, is not significantly different for men and women, and is significantly positively associated with all the health problems included in the model. The age and particularly education effects are steeper in the US than in the Netherlands. The DIF corrected results for the US imply an even larger fall in work disability across education groups, since a higher inital threshold for work disability is used by those in the lowest education category compared to the higher educated groups. In the Netherlands, there is no evidence of a relation between response scales and education level, and the relation between education level and work disability is much weaker, both before and after correcting for DIF.

In the US, we estimate that women use higher thresholds than men. In the Netherlands, there is no evidence that male and female respondents use different thresholds. The effect of gender on work disability is insignificant for both countries. In general, there is substantially less within country variation in thresholds among the more homogeneous Dutch. Since all regressors are included as deviations from their sample means, the significantly negative coefficient on the dummy for NL in γ^1 implies that the respondent with average characteristics in the Netherlands uses a significantly lower first threshold than an otherwise identical respondent in the US.

Pain and emotional problems are more important causes for work disability in the Netherlands. In both countries, respondents with emotional problems use lower thresholds than those without these problems, but the difference is larger in the US. Thus, the effect of emotional problems on work disability is overestimated without correcting for DIF, particularly in the US.

Finally, we discuss some parameter estimates for the benchmark model not presented in Table 3. In addition to the use of different thresholds by gender, we find evidence that the threshold used was lower if a female name was used in the vignette discription instead of a male name: Negative estimates of θ^d were found for all three domains, and were significant for pain and CVD. Thus, for a given vignette description, a male vignette person is seen as more work disabled than a female vignette person, by both male and female respondents.[11]

We find significant unobserved heterogeneity in the thresholds, leading to a positive correlation between several vignette evaluations of the same respondent, and between self-reports and vignette evaluations. The estimated standard deviation of the unobserved heterogeneity term u_i in (1.3) is 3.62 with standard error 0.15. With the standard deviation of unobserved heterogeneity in respondent work disability set to 10 and the estimated standard deviation of the error term in the vignette equation of 5.51, this implies a correlation coefficient of 0.08 between self-reports and vignette evaluations, with observed charactersitics X_i constant.

C. Comparisons of Work Disability Across Countries

In this section we present the results of simulations based on our models for the basic question of how important response scale differences are in explaining differences between the US and The Netherlands in reported rates of work disability. We focus on the 51-64 age group and use sample weights at the respondent level that are provided with the HRS and CentERpanel to make the samples population representative of the 51-64 age groups in the two countries. The simulations take the explanatory variables in the two samples as given and simulate values of work disability and of the thresholds, using US thresholds for both the US and the Dutch sample.

We first consider the model without DIF and the benchmark model, see Table 4. The model without DIF predicts work disability rates of 22.7%in the US and 35.8% in the

Netherlands, a difference of more than 57%. This is similar to what is in the raw data and to what the benchmark model predicts if the Dutch use the Dutch scales and the US respondents use the US scales (numbers not shown). If the Dutch would use the US response scales, however, the simulations show that reported work disability would be reduced to 28.3%, only 23.4% more than the US rate. Thus, for this age group, the difference in response scales explains more than half of the difference in the raw data.

Table 4 also uncovers to what extent, according to the same benchmark model, chronic health conditions explain the work disability rate in both countries, again using the US thresholds for the Dutch as well as the US respondents. Consider an evaluation of the impact of a health condition *j*. Let *P(A)* and *P(B)* be the (predicted) work disability rates in countries A and B and let $P(A)^{-j}$ and $P(B)^{-j}$ be the predicted work disabilities in countries *A* and B for the "counterfactual" situation that nobody would suffer from health problem *j*. $P(A)-P(A)^{-j}$ can then be interpreted as the work disability rate in country *A* due to that health problem and similarly for country B. Note that $P(A)-P(A)^{-j}$ depends both on the prevalence of the health problem and on the sensitivity of the probability of work disability to that health problem (i.e., on the 'impact' effect, driven by the corresponding coefficient in β). We can write:

$$(3.3)\ P(A)-P(A)^{-j} = {}^{1}\!/_{N_A} \sum_{i\in A}\{g(x_i,b_A)-g(x_i^{-j},b_A)\} = [\sum_{i\in A} x_{ij}/N_A][\sum_{i\in A, x_{ij}=1} \Delta g(x_i^{-j},b_A)/\sum_{i\in A} x_{ij}]$$

where $g(x_i,b_A)$ is the probability that an individual with characteristics x_i and parameter vector b_A has a work limitation; x_i^{-j} is the vector x_i with its *j*-th element x_{ij} equal to zero.

The first factor in (3.3) is the fraction in country *A* that suffers from the chosen health problem (the "prevalence effect" for country *A*). In the second term, $\Delta g(x_i,b_A)$ is the marginal effect for a dummy variable, the difference if it is set to 1 or 0, with other variables set to their

values for observation *i*. Thus the second term can be seen as the average impact of the health problem on the probability of work disability for those who have the health problem.

Table 4 shows that prevalence of all chronic health conditions for the age group 51-64 is larger in the US than in the Netherlands, with the exception of pain. Pain has by far the largest impact on work disability of all chronic conditions in both countries. Moreover, the impact of pain is much larger in the Netherlands than in the US. In the model with DIF, it explains a work disability rate of 13.9%-points in the Netherlands and 7.9%-points in the US. All health conditions in the model can jointly explain 20.6%-points of the total work disability in the Netherlands and 16.1%-points in the US. All health conditions combined thus explain 4.5% of the total 5.5% points differential in work disability between the two countries. Pain by itself could however already explain the total difference.

D. Sensitivity Analysis

Table 5 presents results for a number of different model specifications. The first two lines are copied from Table 4, showing that according to the benchmark model, most of the difference in reprted work disability between the two countries is due to response scale differences. The third line presents the results for a model similar to the benchmark model, but ignoring individual unobserved heterogeneity in the scales (u_i in equation (1.3)). The outcomes hardly change. The fourth line of Table 5 presents results for the benchmark model in which the three categories moderate, severe and extreme work disability are combined for the five-point scale self-reports and the vignette evaluations. We see hardly any difference with the benchmark model. This is what we expected, since the *yes* or *no* scale is driven by the distinction between *none*, *mild*, and *moderate*, while the thresholds between the more severe categories play no role.

Another issue for our sensitivity analysis is the use of health conditions as regressors.

Until now, we have ignored potential measurement and cross-country comparability issues with these reported health conditions, which may play a role (Michael Baker et al. (2004)). If there are systematic differences in reporting health conditions across countries, this might bias our results. We therefore re-estimated the benchmark model excluding the health conditions from the equations for work disability and the thresholds. The results in the fifth row of Table 5 show that this makes little difference for the predicted work disability rate. The estimated difference between The Netherlands and the US using US response scales in both countries becomes 17.8%, somewhat lower even than the 23.4% in the benchmark model.

The benchmark model assumes that response scale differences are the same in all domains of work-related disability. That is, if US respondents are harder on people with pain problems than Dutch respondents, then they are also harder on people with emotional problems or people with heart problems. To check whether this assumption is reasonable, we have re-estimated the benchmark model using the vignettes in only one of the three domains. The resulting predictions are presented in the next three rows of Table 5. They show that vignettes in the three domains lead to different conclusions. If we use the affect vignettes only, the correction for response scale differences is very large, and response scale differences explain almost the complete difference in the self-reported rate of work disability. But if we only use the vignettes on heart problems, the opposite conclusion is obtained: respondents in the US and The Netherlands use similar response scales, and less than a quarter of the difference in self-reported work disability rates is explained by response scale differences. For the pain vignettes, the results are in between these two extremes and similar to those for the benchmark model. Thus these results cast doubt on the assumption of common response scales across health domains.

The final row presents results from a more general model, which accounts for different response scales for the three domains. This model is explained in detail in Appendix 3 (available on line). Although the assumptions underlying this model are very different from those of the benchmark model, the resulting estimate of the percentage difference in work disability between the U.S. and The Netherlands is almost identical to the estimate for the benchmark model.

IV. Conclusions

In sharp contrast to the believed similarity in their health outcomes, workers in different western countries report very different rates of work disability**.** This contradiction continues to be seen as a major unresolved puzzle. In this paper, using new data from the US and The Netherlands, we offer a partial resolution of the puzzle. We find that a large part of observed differences in reported work disability stems from the fact that residents of the two countries use different response scales in answering standard questions on whether they have a work disability. Essentially for the same level of actual work disability, Dutch respondents have a lower response threshold in claiming disability than American respondents do.

We reached this conclusion by implementing a vignette methodology into Internet surveys in both countries. Our vignettes gave respondents the same simple scenarios in which hypothetical workers varied in the objective circumstances of their work disability. Respondents were asked to rate the extent of that disability. Especially in the more subjective health domains of pain and emotion, the evidence is strong that American respondents use a 'tougher' standard when assigning a work disability status. While explaining these different standards is an important research question in itself, there seems little question that they exist. While one may quarrel with the specific assumptions in each modelling approach in the paper, the similarity of their implications for explaining international differences in work disability is striking.

In addition to their role in explaining across country differences, vignettes are a useful tool in helping us understand within country differences in reporting. Using vignettes given to Americans show that different thresholds are used by three of the most widely used empirical determinants of work disabilty—sex, education, and age. Such differences also have implications for the use of self-reported health as an explanatory variable in models explaining, e.g., labor force participation or mortality. While self-reported health is typically strongly correlated with objective health indicators (see Ellen Idler and Yael Benyamini, 1997), its measurement scale may well vary systematically with other explanatory variables, biasing the estimated coefficients of these other variables. For example Hendrik Juerges (2005) finds that next to self reported health, socio-economic variables have an independent predictive effect on mortality. Without further information one cannot determine whether the effect of the socio-economic variables is real or whether it simply picks up systematic differences in the self reported health measures. Vignettes provide an opportunity to directly analyze scale differences and correct for them.

Vignettes represent a potentially important new methodological tool that may aid in the analyses of other applications besides health and disability. Anytime threshold scales categorize individual responses, the question will arise on whether people really differ or whether they are simply not using the same scales. Vignettes can help answer that question in such varied applications as general well being-scales, political efficacy (King et al. 2004), health problems, consumer satisfaction, risk measurement, and perception of poverty. The application of a new technique like vignettes also poses new methodological questions. Internet surveys appear to be a powerful tool to address such questions.

REFERENCES

Aarts, Leo, Richard Burkhauser and Phillip De Jong. "Curing the Dutch Disease: An International Perspective on Disability Policy Reform." Aldershot, Avebury. 1996.

Autor, David and Mark Duggan. "The Rise in the Disability Rolls and the Decline in Unemployment." *Quarterly Journal of Economics,* 2003, 118(1):157-206.

Baker, Michael, Mark Stabile and Catherine Deri. "What Do Self-reported, Objective Measures of Health Measure?" *Journal of Human Resources*, 2004, 39(4):1094-1115.

Bound, John and Richard Burkhauser. "Economic Analysis of Transfer Programs Targeted on People with Disabilities." *Handbook of Labor Economics, Vol. 3C*, Orley Ashenfelter and David Card (eds.), 1999, 3417-3528.

Bound, John. "Self-reported versus Objective Measures of Health in Retirement Models." *Journal of Human Resources,* 1991, 26(1):106-138.

Burkhauser, Richard and Mary Daly. "Policy Watch: U.S. Disability Policy in a Changing Environment." *Journal of Economic Perspectives,* 2002, 16(1):213-224.

Burkhauser, Richard, Mary Daly, Andrew Houtenville and Nigar Nargis. "Self-Reported Work Limitation Data—What They Can and Cannot Tell Us." *Demography*, 2002, 39(3):541-555.

Currie, Janet and Brigette Madrian. "Health, Health Insurance and the Labor Market." *Handbook of Labor Economics,* 1999*, Vol. 3C*, O. Ashenfelter and D. Card (eds.), 3309-3416.

DeLeire, Thomas. "The Wage and Employment Effects of the American with Disabilities Act." *Journal of Human Resources*, 2000, 35(4):693-715.

Haveman, Robert, and Barbara Wolfe. "The Economics of Disability and Disability Policy." *Handbook of Health Economics,* 2000*,* Vol. 1B, J. Newhouse and A. Culyer (eds.), North Holland, Amsterdam, 2000, 995-1051.

Idler, Ellen L. and Yael Benyamini. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." *Journal of Health and Social Behavior*, 1997, 1:21-37.

Juerges, Hendrik. "Self-assessed Health, Reference Levels, and Mortality." Working Paper, Mannheim Institute for the Economics of Aging, 2005.

Kerkhofs, Marcel and Martin Lindeboom. "Subjective Health Measures and State Department Reporting Errors." *Health Economics*, 1995, 4:221-235.

King, Gary, Christopher Murray, Joshua Salomon, and Ajay Tandon. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review,* 2004*,* 98(1), 567-583.

Salomon, Joshua, A. Tandon, and Christopher Murray. "Comparability of Self rated Health: Cross Sectional Multi-country Survey Using Anchoring Vignettes." *British Medical Journal,* 2004, 328 (7434), 258-260.

Stapleton, David and Richard Burkhauser (eds.). "The Decline in Employment of People with Disabilities: A Policy Puzzle." Kalamazoo, MI: W.E. UpJohn Institute for Employment Research, 2003.

ENDNOTES

∗ We are gateful to the editor and three anonymous referees for useful comments.

[1] Vignette questions have been applied successfully in recent work on international comparisons of health and political efficacy (King et al., 2004; Joshua Salomon et al., 2004).

[2] The γ^j for $j>1$ will still be identified.

[3] The gender of each vignette person was randomly assigned.

[4] http://cdata4.uvt.nl/websitefiles/representativiteit2005.pdf provides a comparison of sample characteristics with the population distribution (in Dutch). Generally the differences are minor; the most important differences being in the domains of age and education: 13.8% of the population is over 65, versus 10.9% in the sample; 62.7% of the population has at least a high school degree, versus 67.3% in the sample.

[5] All questions about health conditions except that on pain are of the form has the doctor ever told *you that you have* …. The distribution of the number of doctor visits is almost the same in both countries, suggesting that this does not lead to a systematic difference.

[6] The full set of vignettes are in Appendix 1 (available on line).

[7] In contrast to the CentERpanel, the RAND MS panel only has respondents with prior Internet access. The critical issue here is not whether the samples we use differ from the population of interest in observable characteristics, since we use sampling weights (and the HRS). The issue is whether those in Internet samples, conditional on observables, use different thresholds and thus give different vignette evaluations. For part of the Dutch sample we know if respondents had Internet access before they joined the panel. We reestimated the equations for the vignette

evaluations with additional dummies for whether a respondent had prior Internet access or not. In all three domains, the coefficients of these dummies turned out to be insignificant. This suggests that our results for the US are not selective due to prior Internet access.

[8] We use 1978 observations from CentERpanel, 672 observations from the RAND MS Internet panel, and 15,740 observations on persons aged less than 75 from HRS 1998.

[9] In addition, we split up the error terms in the work disability equation in a part that is common to the two-point and the five-point report, and parts that are idiosyncratic for the two-point and five-point reports. The former can be seen as unobserved heterogeneity, the latter as noise. The standard deviation of the heterogeneity term is normalized to 10.

[10] Appendix 2 (available on line) has the complete estimation results for the benchmark model.

[11] The interactions of gender of the vignette person and gender of the respondent were insignificant.

Table 1. % With Work Disability and Health Conditions by Age—US and Netherlands

	Age Group			
	25-34	**35-44**	**45-54**	**55-64**
United States	7.4	11.3	17.6	25.9
Netherlands	17.3	22.8	36.8	37.1
Health Conditions				
United States				
Diabetes	2.0	3.8	6.8	10.8
Arthritis	3.8	9.3	19.1	28.8
Hypertension	7.3	11.6	21.9	35.4
Heart problem	1.0	2.3	5.3	11.9
Stroke	0.5	0.7	1.5	3.9
Emotional problems	4.7	5.3	6.6	6.0
Netherlands				
Diabetes	1.2	2.1	3.3	6.5
Arthritis	3.7	8.7	11.1	16.5
Hypertension	6.6	13.1	22.2	29.9
Heart problem	0.4	2.0	3.9	12.9
Stroke	0.2	0.5	0.8	2.7
Emotional problems	13.2	11.7	14.8	10.0

US data are from PSID. Netherlands data are from CentERpanel. All data are weighted.

Table 2. Vignette Evaluations in United States and Netherlands

Pain vignettes	**1**		**2**		**3**		**4**		**5**	
Limited?	**NL**	**US**	**NL**	**US**	**NL**	**US**	**NL**	**US**	**NL**	**US**
Not at all	24.9	38.7	10.5	30.6	0.4	0.2	0.5	0.5	0.5	0.2
Mildly	63.3	48.9	53.5	46.4	6.2	7.3	11.9	8.6	7.3	2.6
Moderately	10.5	10.9	29.4	21.1	26.6	30.7	33.8	38.5	31.1	15.4
Severely	1.3	0.5	6.3	1.0	50.9	47.1	43.9	39.9	46.3	58.3
Extremely	0.1	1.0	0.3	0.9	16.0	14.7	9.9	12.4	14.9	23.5
Affect vignettes	**1**		**2**		**3**		**4**		**5**	
Limited?	**NL**	**US**	**NL**	**US**	**NL**	**US**	**NL**	**US**	**NL**	**US**
Not at all	96.8	97.7	32.2	55.1	12.4	34.2	7.4	23.0	1.3	8.4
Mildly	2.4	0.9	54.0	34.1	43.6	38.4	35.3	37.9	5.4	11.2
Moderately	0.5	0.4	11.8	8.7	31.5	21.3	39.7	29.1	14.8	20.1
Severely	0.3	0.2	1.8	1.2	11.8	5.8	16.2	8.7	43.3	42.9
Extremely	0.1	0.9	0.2	0.9	0.8	0.4	1.4	1.2	35.3	17.3
CVD vignettes	**1**		**2**		**3**		**4**		**5**	
Limited?	**NL**	**US**	**NL**	**US**	**NL**	**US**	**NL**	**US**	**NL**	**US**
Not at all	88.8	94.1	20.5	26.7	9.1	12.9	7.1	7.5	1.9	3.3
Mildly	9.8	4.9	43.3	31.9	49.1	35.7	36.6	21.2	18.6	15.0
Moderately	1.0	0.2	26.2	27.4	28.7	32.7	31.6	32.4	36.3	32.5
Severely	0.4	0.0	9.7	12.4	12.2	16.5	20.8	30.1	34.3	39.1
Extremely	0.1	0.9	0.4	1.7	0.9	2.3	3.8	8.9	8.9	10.3

Sources: Netherlands: CentERpanel, August 2003, 1978 observations; US: RAND MS Internet Panel, January 2004, 672 observations. See Table A2 in the appendix for all vignette descriptions.

Table 3. Selected Estimates Benchmark Model

	Model without DIF		Complete Model			
	Work disability		First Threshold Parameter		Work disability	
	β	s.e.	γ^1	s.e.	β	s.e.
Constant	-6.71	0.46*	0.00	0.00	-7.91	0.36*
Ed med	-3.19	0.30*	-0.46	0.39	-3.77	0.49*
Ed high	-5.32	0.42*	-1.02	0.46*	-6.56	0.62*
Age/100	78.61	25.76*	49.53	9.74*	150.59	26.91*
(Age/100)^2	-46.71	20.47*	-49.18	8.50*	-112.87	21.42*
Female	-0.28	0.26	1.04	0.32*	0.61	0.42
Hypertension	2.04	0.27*	0.41	0.34*	2.55	0.43*
Diabetes	4.03	0.36*	-1.58	0.55*	2.66	0.63*
Cancer	2.63	0.41*	0.23	0.50	2.97	0.64*
Disease of lung	5.95	0.44*	0.00	0.65	6.19	0.77*
Heart problem	5.64	0.35*	-0.19	0.62	5.88	0.67*
Emotion	6.18	0.39*	-1.99	0.53*	4.67	0.64*
Pain	10.62	0.39*	-0.64	0.33	10.92	0.51*
Interactions with dummy NL						
Constant	4.16	0.56	-3.05	0.27*	1.94	0.60*
Ed med	2.81	0.86*	0.34	0.44	3.12	0.96*
Ed high	2.46	0.94*	1.38	0.51*	3.94	1.05*
Age/100	-26.97	29.63	-51.09	10.09*	-102.40	30.57*
(Age/100)^2	5.17	24.78	50.05	8.67*	74.55	25.57*
Female	1.28	0.73	-1.07	0.36*	0.41	0.81
Hypertension	-1.29	0.87	-0.32	0.39	-1.79	0.95
Diabetes	1.52	1.60	0.68	0.76	1.94	1.73
Cancer	-0.28	1.49	-0.23	0.83	-0.43	1.59
Disease of lung	0.92	1.33	-0.05	0.80	0.02	1.52
Heart problem	2.88	1.25*	0.39	0.71	2.63	1.43
Emotion	1.88	0.99	1.14	0.59	2.36	1.16*
Pain	4.83	0.84*	0.95	0.40*	4.23	0.93*

Notes : Normalization: $\sigma_r^2 = 100$; * : significant at two-sided 5% level.
All regressors defined in deviations from their overall means.

Table 4. Predicted Work Disability Age Group 51-64—US versus NL

Panel 1

	Percentage Work Disabled		
	NL	**US**	**% Difference NL-US**
Model without DIF	35.8	22.7	57.5%
Benchmark model with DIF	28.3	22.8	23.4%

Panel 2

Role of Chronic Health Conditions, Benchmark Model

	prevalence		impact effect on prob. disab.		contribution to work disability		
	NL	US	NL	US	NL	US	NL-US
high blood	28.26	38.47	1.73	5.48	0.49	2.11	-1.62
diabetes	5.45	11.43	11.14	6.55	0.61	0.75	-0.14
cancer	5.45	7.37	5.67	6.58	0.31	0.48	-0.18
lung	5.94	7.16	14.65	15.35	0.87	1.10	-0.23
heart	12.15	13.68	20.90	14.54	2.54	1.99	0.55
emotional	10.50	14.26	17.33	12.05	1.82	1.72	0.10
pain	33.66	27.55	41.34	28.70	13.91	7.91	6.01
all health conditions					20.55	16.06	4.49

Notes: CentERpanel for The Netherlands and HRS 1998 for the US, weighted with sampling weights.

Table 5. Predicted Work Disability Age Group 51-64 using US Response Scales—Several Models

	Percentage Work Disabled		
	NL	**US**	**%Difference NL-US**
Model without DIF	35.8	22.7	57.5%
Benchmark model using all vignettes	28.3	22.8	23.4%
No unobserved heterogeneity in thesholds	28.0	22.7	23.4%
Model combining moderate, severe, extreme	28.3	22.8	24.4%
Model not using health conditions	27.2	23.1	17.8%
Model using affect vignettes only	24.3	22.8	6.8%
Model using pain vignettes only	28.2	22.8	23.9%
Model using cvd vignettes only	32.9	22.8	44.3%
Model with non-common thresholds	28.5	23.1	23.6%

Note: CentERpanel and HRS; weighted using sample weights at respondent level. Predicted work disability on two-point scale.

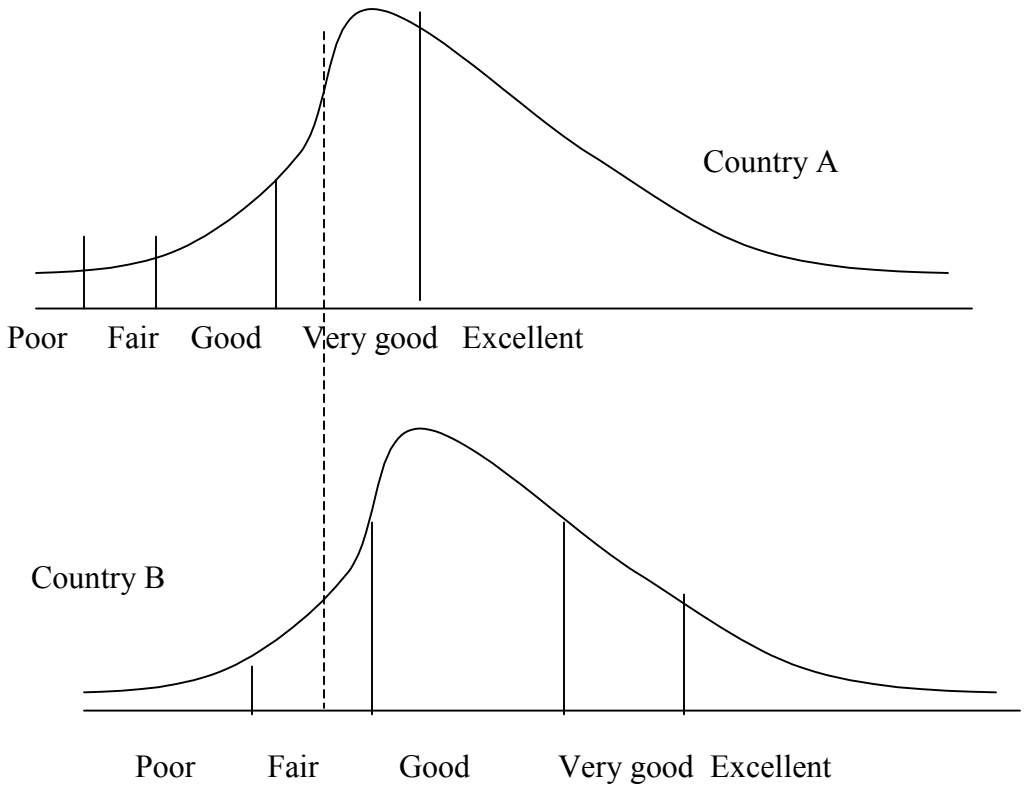

Figure 1: Comparing self-reported health across two countries in case of DIF