Self-reported Work Disability in the US and The Netherlands

Arie Kapteyn, James P. Smith, RAND
Arthur van Soest, RAND & Tilburg University

November 2004

Abstract

Self-reported work disability is analyzed in the US and The Netherlands. The raw data show that Dutch respondents much more often report that they have a work limiting health problem than respondents in the US. The difference remains when controlling for demographic characteristics and observed onsets of health problems. Respondent evaluations of work limitations of hypothetical persons described in vignettes are used to identify the extent to which the differences in self-reports between countries or socio-economic groups are due to systematic variation in the response scales. A model that assumes the same response scales for different health domains is compared with a model that allows for domain specific response scales. Results of both models suggest that about half of the difference between the self-reported rates of work disability in the US and The Netherlands can be explained by response scale differences.

JEL codes: J28, I12, C81
Keywords: Work limiting disability, Vignettes, Reporting bias

This research was supported by grants from the National Institute on Aging and the Lasker Foundation to the National Bureau for Economic Research, by a grant from the National Institute on Aging to RAND. The paper benefits from useful comments at the NBER Disability group meeting in Charleston SC and Jackson Hole WY, the RTN meeting in Edesheim, the NBER Summer School, the PAA meetings, and seminars in Amsterdam, Dublin, Utrecht, Tilburg, and Bristol.

1. Introduction

Reducing work disability among the working population is an important issue on the scientific and policy agenda in many industrialized countries. See, for example, Haveman and Wolfe, 2000, or Bound and Burkhauser, 1999. The fraction of workers drawing some form of disability benefit 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 contribute to explaining the differences in disability rolls (see, e.g., Bound and Burkhauser, 1999, Burkhauser and Daly, 2002, and 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 rates of work disability reflect true differences across countries in disabilities affecting work. But if response scales differ systematically, adjustments for this must be made before conclusions about international differences in true work disability can be drawn. The problem is similar to the problem of systematic reporting differences across socio-economic groups. See, e.g., Bound (1991), Currie and Madrian (1999), Kerkhofs and Lindeboom (1995), and Burkhauser et al. (2002).

Disability is an important program in many countries, and one that until recently was growing rapidly over time. The number of people on disability programs is substantial, particularly among men and women in the age groups 45-64. For the US, Autor and Duggan (2003) find that the numbers of disability insurance (DI) recipients per 1000 men and women in the age group 55-64 have increased from 96 to 108 (men) and from 43 to 72 (women) between 1984 and 1999. Bound and Burkhauser (1999) report that in 1995, the number of DI recipients per 1000 workers was 103 in the age group 45-59 and 314 in the age group 60-64. Both numbers have grown substantially in the early nineties. There are also substantial differences amongst OECD countries. For example, the numbers of DI recipients per 1000 workers in the age category 45-59 were 87 for Germany and 271 in The Netherlands. According to Eurostat (2001), the number of 16-64 year olds receiving disability and sickness benefits is less than 3% in Italy and Greece, but almost 10% in Denmark and more than 12% in the UK.

The paper puts forth a new approach to the measurement of work disability. In particular, we utilize a vignette methodology to evaluate how people within and across different 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 also evaluate themselves the severity of work disability problems of hypothetical scenarios and people. Vignette questions have been applied successfully in recent work on international comparisons of health and political efficacy (King et al., 2004; Salomon et al., 2004).

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 and eligibility for government programs that provide income support for people with a work disability, and perhaps national norms about the appropriateness of not working when work disabled (see, e.g., Aarts, Burkhauser, and De Jong, 1996). However, given their similar levels of economic development and access to modern medical technology and treatment, 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 some of the most salient research issues that have dominated the scientific literature on work disability.

A unique aspect of this research is that we are able to address the issues in a classic random experimental form. This is because we have access to Internet samples in both countries allowing us to randomly place experimental disability modules into these panels. These samples are the Dutch CentERpanel for The Netherlands and the RAND MS Internet panel for the United States, both of which are described in detail in the next section

The remainder of this paper is organized as follows. In the next section, we describe our data, discuss some measurement issues and present descriptive statistics on self-reported work limitations in the US and The Netherlands. In Section 3 we present some illustrative estimates of the determinants of work disability estimated across the countries of interest. These models do not correct for the possibility that respondents in The Netherlands and the US may apply different scales when responding to questions about work limitations. Section 4 describes the vignette methodology and two different models that can be used to correct for scale differences

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

across countries. Section 5 presents the empirical results for these models and some variants. Conclusions follow in Section 6. Some details of specifications, survey questions, and results are presented in the appendix.

2. Data Sources and Work Disability Prevalence

In this research, we use information obtained from two Internet surveys, which we conducted in both countries, combined with the Health and Retirement Study (HRS). For The Netherlands, we used the Dutch CentERpanel, which includes about 2,250 households who have agreed to respond to a set of 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 already have Internet access, they are provided with Internet access (and if necessary, a set-top box). Thus, the CentERpanel is representative of the Dutch population except the institutionalised. The sample that we use to estimate our models consists of about 2,000 respondents who participated in several interviews with questions on work disability in 2003.

From multiple waves of the data that have been collected in the past, the CentERpanel has a rich set of variables on background and demographic characteristics of the respondent and household, their income and labor market status, and several salient dimensions of health. In August 2003, we collected work disability self-reports and vignette evaluations (described below) in the CentERpanel. The Internet infrastructure makes the CentERpanel an extremely valuable tool to conduct experiments, with possibilities for randomization of content, wording, question and response order, and regular revisions of the design. Production lags are very short, with about one month between module design and data delivery. For example, based upon our initial analysis, we fielded a second wave in October with different wordings of the vignette questions. A third wave of experiments was administered in December 2003.

The RAND MS Internet panel has been recruited from respondents of age 40 and older to the Monthly Survey (MS) of the University of Michigan's Survey Research Center (SRC). The MS is the leading consumer sentiments survey that incorporates the long-standing Survey of Consumer Attitudes (SCA) and produces the widely used Index of Consumer Expectations. SRC asks MS-respondents age 40 or older if they have Internet access and, if yes, whether they would be willing to participate in Internet surveys. Those who agree to participate are added to the panel of households to be interviewed regularly over the Internet. The sample that we use for estimation consists of 672 respondents. Ultimately, the sample will be extended to 1,000 Internet respondents.[2] Because of the relatively small sample 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 large and representative cohort interviewed when at age 51-61. The HRS sample has self-reports on work disability like the RAND MS Internet survey and the CentERpanel, but does not have vignette questions.

2.1 Question differences and prevalence

One of the difficulties in making international comparisons is that the form and wording of questions about work disability differ across countries and even within countries in different surveys. Question wording is often thought to be a possible source of differences across and within countries (Stapleton and Burkhauser, 2003)**.** To test the impact of question wording, we randomly assigned the disability questions contained in the National Health Interview Survey (NHIS), Current Population Survey (CPS), and the Health and Retirement Survey (HRS) to our Internet respondents[3]. None of the variants appeared to matter for the probability of describing oneself as having a work disability. What does matter though is whether the scale used to evaluate work disability is a two point yes/no commonly used in the United States or a more graded five-point scale typically used in European countries including The Netherlands.

The question we use 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?"

Respondents in the US survey answer on a two-point scale (*yes* or *no*) while the possible answers are arrayed on the following 5-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, and (5) yes, I am extremely limited—I cannot work.

[2] By the same mechanism a control group is drawn of respondents who do not necessarily have Internet access and are interviewed by phone. Ultimately this control group will comprise 500 respondents. The number of available phone interviews is currently 225 and these observations are not used in the analysis, to avoid contamination by possible mode effects between phone and Internet interviews.

[3] To be precise, the three different wordings were randomly assigned to respondents in yet another Internet survey : respondents to the 2002 HRS, who have Internet access and agreed to participate in an Internet survey in 2003. The survey covered about 2500 respondents.

To evaluate the impact of the alternative scale, randomly half of CentERpanel respondents in the second wave of our vignette experiments were given the disability question on a yes/no scale as in the US. Given that the first two waves of our experiments were only a few months apart so that disability reports should not change that much, for these respondents one can compare the answers to this question to that given on the 5-point scale a few months earlier.

The results are presented in Table 1. For all but one row in the 5-point scale, the correspondence is remarkably close. Ninety-six percent of those who answered they were not at all disabled on the 5-point scale also said that they were not when using the HRS dichotomous scale. Similarly, more than 90% of Dutch respondents who said that they were more than somewhat limited replied that they had a work disability on the two-point scale.

The ambiguity occurs within the somewhat limited category, which splits about 50/50 when offered an opportunity to simply respond yes or no. These are people who are clearly on the margin in terms of their work disability problems. When offered a stark yes or no choice, some will resist disability labeling. But if given a more nuanced set of alternatives, they report some degree of disability.

Table 1

Correspondence Between 5- and 2-point Scale in Dutch Panel

5-point scale work limitations	% in 5-point category	% disabled on 2-point scale in each 5-point category
not at all	61.8	4.3
somewhat limited	22.5	56.1
rather limited	9.9	91.2
severely limited	2.2	93.1
very severely limited	3.6	92.1

Source: Dutch CentERpanel.

Table 2 shows reported US disability rates by age from the PSID and Dutch disability rates obtained from CentERpanel using the same two-point scale. Especially for middle age workers—say those between ages 45-64—Dutch rates of reported work disability are about 15 percentage points higher than those in the United States even when the same question is asked in both countries. We will turn to explanations for this difference in later sections. The final row in Table 2 shows work disability in the Netherlands derived from the 5-point scale, defining everyone who reports a mild limitation or worse as work disabled. As expected from Table 1, this gives even higher work disability rates for all age groups.

Table 2
% With Work Disability by Age—US and Netherlands

	Age Group				
	25-34	35-44	45-54	55-64	65+
US, 2 point scale	7.4	11.3	17.6	25.9	38.8
Netherlands					
2-point scale*	17.2	23.6	38.7	37.4	38.8
5-point scale	25.7	30.3	42.7	44.2	53.6

US data are from PSID. Netherlands data are from CentERpanel. All data are weighted.
*Derived from five-point scale: anyone reporting mildly limited or worse is considered work disabled.

Reporting differences on health status between the two countries are not limited to the domain of work disability. Table 3 lists respondents' evaluation of their health along the familiar 5-point scale—excellent, very good, good, fair, and poor. Since this comparison involves two populations where as a first approximation their 'true' health status is unlikely to be very different, it is apparent that the Dutch and Americans use very different criteria to place themselves in these five categories.

Table 3
Comparison of Self-reported General Health Status

	Netherlands	US
Excellent	5.8	24.7
Very Good	23.9	36.0
Good	56.2	28.1
Fair	11.8	8.9
Poor	1.1	2.3

US data are from PSID. Netherlands data are from CentERpanel. Ages 25-64 in both countries. All data are weighted.

The circumspect Dutch appear to run to the center, not willing to make health claims at either the top or bottom while the ever optimistic Americans are four times more likely to state that they are in excellent health. While the data in Table 3 refer to general health status, we shall see below that this general tendency of the Dutch to avoid the extremes in self-categorization will have important consequences for their reported levels of work disability as well.

3. Comparisons of Work Disability Probits Across Countries

The principal question that we ask in this paper is how much of the reported differences among these countries reflect differences in response scales and how much reflects actual differences in true work disability. Our first step in that inquiry is to estimate standard models for self reported work disability in both countries as a function of an also standard set of demographics and health. Given the concentration of work disability rates during the preretirement years, the models for both countries are based on data for the age range 51-64. The starting age of 51 is determined by the age cut-off in the HRS.[4]

All models are probits estimating the probability that a respondent reported having a work disability. The Dutch models are estimated using the same 2-point scale variable as in the HRS. The covariates in the model include education, gender and the following health attributes—whether one has hypertension, diabetes, cancer, disease of the lung, heart disease, stroke, arthritis, emotional problems or suffers from pain.[5] The estimates for both countries are given in Table A1 in the appendix.

Our goal with these models is twofold—to uncover the principal factors that led to a report of work disability and to isolate the sources of the international difference in reported work disability. To see how we accomplish this goal, consider for example an evaluation of the impact of a single health condition j. Let $P(A)$ and $P(B)$ be the (predicted) work disability rates in country A and country B (for a given age group) and let $P(A)^{-j}$ and $P(B)^{-j}$ the predicted work disabilities in country 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 this assignment of importance to this health condition 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 corresponding coefficients in β_A); we will separate these two below.

The difference in work disabilities in the two countries can be expressed using the following decomposition:

[4] We also estimated these models on the PSID and the full age range of the Dutch sample and the results were qualitatively similar to those reported here.

[5] Self-reported general health status is available in both countries, but Table 3 suggest that the Dutch and Americans use different criteria to place themselves within the five categories so that the self-reported health measures in the two countries are not comparable. We therefore do not include self-reported general health status as a regressor.

$$P(B) - P(A) = [P(B)^{-j} - P(A)^{-j}] + [P(B) - P(B)^{-j}] - [P(A) - P(A)^{-j}] \qquad (3.1)$$

The first term on the right hand side can be interpreted as the difference between work disability prevalence in the two countries that is *not* due to the chosen health problem. The sum of the second and third term is then the part that is due to the chosen health condition. The latter two terms can be further separated in a 'prevalence' effect (the percentage with the health problem) and an 'impact ' effect (the impact of the health problem on work disability). We can write:

$$\begin{aligned} P(A) - P(A)^{-j} &= {}^{1}\!/_{N_A} \sum_{i \in A} \{g(x_i, b_A) - g(x_i^{-j}, b_A)\} = \\ &[\sum_{i \in A} x_{ij} / N_A][\sum_{i \in A, x_{ij}=1} \Delta g(x_i^{-j}, b_A) / \sum_{i \in A} x_{ij}] \end{aligned} \qquad (3.2)$$

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 *i*-th element x_{ij} equal to zero.

The first factor is the fraction in country *A* that suffers from the chosen health problem (the "quantity effect" for country *A*). In the second term, $\Delta g(x_i, b_A)$ is the marginal effect ("partial derivative") 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 marginal effect for those who have the health problem.

The same decomposition can be used for all co-variates in the model (both health and non-health dummy variables) allowing us to compare the importance of each to the reported rates of work disability in each country and the difference between countries. Table 4.A lists the estimated contribution of each factor for the age group 51-64 in The Netherlands while Table 4.B does the same for the US. Table 5 presents a summary of the relative contributions of different sets of factors toward explaining the differences between the two countries in reported rates of work disability. For this relative asssement, for reasons that will soon become apparent we divide covariates into six groups—heart problems and stroke, the other so called 'objective' health factors (hypertension, diabetes, cancer, diseases of the lung), arthritis, pain, emotional problems, and demographics (education and gender).

Consider first the other 'objective' health conditions other than heart problems . As summarized in the third columns of Tables 4.A and 4.B, prevalence rates of these conditions are actually higher in the United States than in The Netherlands. Estimated marginal effects of having these conditions on the work disability rate are larger in the US in some cases (e.g., diabetes) and smaller in other cases (e.g., lung disease). Collectively, these health conditions would imply a slightly higher rate of work disability in the United States, as summarized in Table 5.[6]

Table 4.A
Decomposition of Dutch Disability—Ages51-64

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	0.27	28.52	0.95
Diabetes	0.15	5.68	2.56
Cancer	0.09	5.49	1.63
Disease of lung	1.63	5.98	27.26
heart problem	1.68	10.55	15.88
Stroke	0.61	2.02	29.96
Arthritis	2.23	14.97	14.89
Emotion	3.33	10.94	30.48
Pain	15.49	33.20	46.65
Female	1.19	53.43	2.23
Ed low	0.71	48.80	1.45
Ed med	0.31	31.05	2.68
Work disability			
Prevalence in sample	0.34		

See Appendix, Table A1 for parameter estimates of underlying probit model.

Table 4.B
Decomposition of US Work Disability—Ages 51-64, HRS

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
Hypertension	1.42	37.96	3.75
Diabetes	1.17	11.23	10.40
Cancer	0.43	7.11	6.00
Disease of lung	0.83	6.88	12.13
Heart problem	1.81	13.16	13.73
Stroke	0.74	3.25	22.68
Arthritis	3.97	42.96	9.25
Emotion	2.50	14.42	17.32
Pain	7.66	27.82	27.53
Female	-0.86	52.72	-1.63
Ed low	2.42	20.96	11.56
Ed med	2.03	56.68	3.59
Work disability			
Prevalence in sample	0.25		

See Appendix, Table A1 for parameter estimates of underlying probit model.

[6] Reporting of such conditions may also be different across nations due to differential physician contact or because the precise criteria for thresholds for medical diagnosis may not be the same.

The one ‘objective’ health condition we separate out in Table 5 is heart disease since we will use it below in our vignettes as a prototype of these health conditions. While the overall effects are small, heart disease also has a higher prevalence in the United States and thus would imply a slightly higher rate of work disability in the United States. It is not central to our argument that the Dutch sample appears healthier than the American one; the main point is that differential levels of these objective health measures are unlikely to account for the much higher self reported work disability rates observed among the Dutch compared to the Americans. Similarly emotion and arthritis appear to be incapable of explaining major differences in work disability between The Netherlands and the US.

This brings us to the one condition that seems a more promising candidate for why disability rates differ between the two countries. In contrast to the other more ‘objective’ health conditions, pain actually has a substantially higher prevalence in The Netherlands compared to the US. The summary in Table 5 singles out pain in particular as a potentially important source of the international difference.

Pain not only has a higher prevalence in The Netherlands, but our probit estimates indicate that pain and emotional problems are among the strongest predictors of work disability. Since these two conditions are more subjective and the more difficult to diagnose, this may indicate that the source of the international differences in reports of work disability rests in these two conditions. It may be that for the same level of pain the Dutch are more likely to say that it constitutes a work disability than are the Americans. This speculation about these possible international differences in reporting leads us to try to test these ideas. Our tests will exploit vignettes on work disability.

To understand the impact of the demographic differences in the two samples—gender and schooling—it should be noted that the choice of benchmark group matters for the decomposition. The highest education level is chosen as the benchmark, since this gives the lowest work disability probability. There is a considerably larger gradient with education in reported work disability in the United States than in The Netherlands, so that the differences in reported work disability between the two countries are largest amongst the most educated. If everyone’s work disability probability would be determined in the way work disability is driven for the higher educated, work disability in the US would be 4.45%-points lower, compared to 1.54%-points in the Netherlands. Gender is significant in the US only, where females are less likely to report a work disability than males. In the Netherlands, the gender dummy has the opposite sign but is insignificant. Combining education and gender, Table 5 shows that if everyone's work disability would be determined as it is for high educated males, reported work disability would be 2.82%-points lower in the US and 2.73%-points in the Netherlands. On the other hand, if reported work disability would be determined as it is for high educated females, it would be 5.22%-points lower in the US and 0.50%-points in The Netherlands.

Table 5
Differences in Dutch and American Disability

Variables	Dutch	American	Dutch-American
'Objective' health conditions	2.14	3.85	-1.71
Heart problems	2.29	2.55	-0.26
Emotion	3.33	2.50	0.83
Arthritis	2.23	3.97	-1.74
Pain	15.49	7.66	7.83
Demographics*	2.73	2.82	-0.09

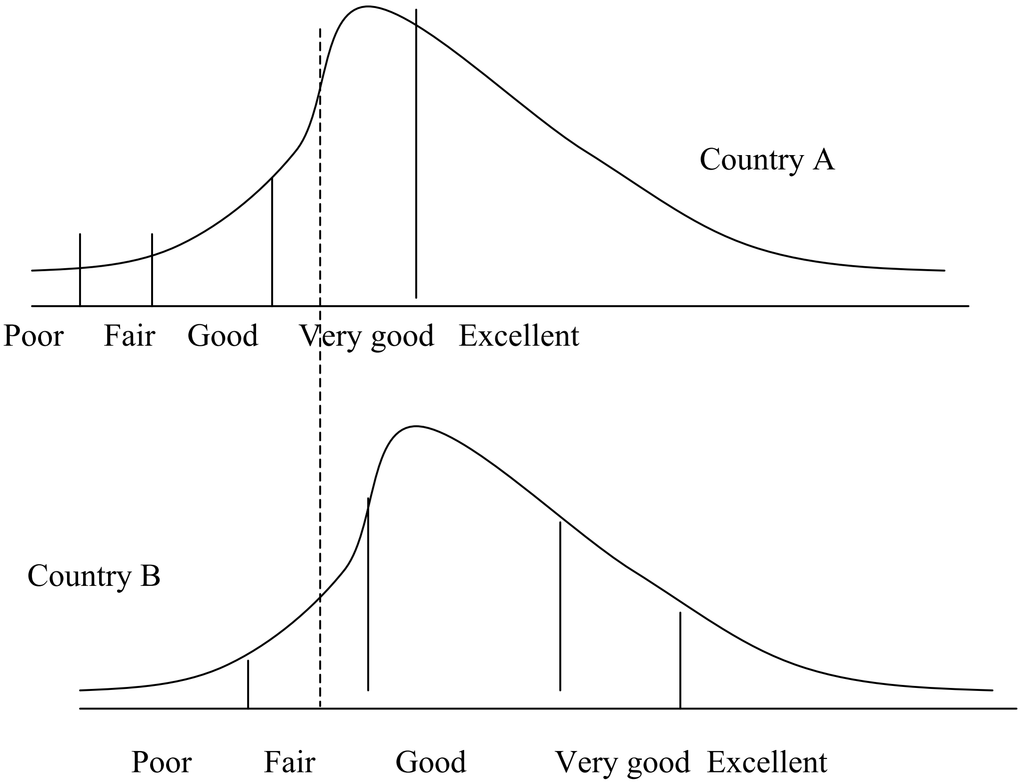
*Benchmark: high educated males.

4. Vignettes

4.1 The Intuition about Vignettes

In this section, we first 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 the example in the figure, people in country A have a much more positive view on a given health status than people in country B. Someone in country A with the 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 the 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 the differences in the response scales (DIF, “differential item functioning,” in the terminology of King et al., 2004) is essential to compare the actual health distributions in the two countries.

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



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 the health of that person on the same five-point scale that was used for the self-report of their own health. Since the vignette descriptions are the same in the two countries, the 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 the actual health is the same in the two countries, the difference in the country evaluations must be due to DIF.

Vignette evaluations thus help to identify differences between the response scales. Using the scales in one of the two countries as the benchmark, the distribution of evaluations in the other country can be adjusted by evaluating them on the benchmark scale. The corrected distribution of the evaluations can then be compared to that in the benchmark country—they are now on the same scale. In the example in the figure, this will lead to the correct conclusion that people in country B are healthier than those in country A, on average. The underlying assumption is *response consistency:* a given respondent uses the same scale for the self-reports and the vignette evaluations. King et al. (2004) provide evidence supporting this assumption by comparing self-reports and vignette evaluations of vision 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 downward about what constitutes good health, 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.

4.2 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 ($Y_{ri,}$ *i* indicates respondent *i*) to the question already discussed in Section 2.1, with answers on a 2-point and a 5-point scale:

"Do you have any impairment or health problem that limits the kind or amount of paid work you can do?"

The questions on work limitations of the vignette persons have the same 5-point scale answering categories and are formulated in the same way ("Does Mr/Mrs X have any impairment or health problem that limits the type or amount of work that he/she can do?"). The answers will be denoted by Y_{li} where each respondent *i* evaluates *L* vignettes *l=1,…,L.*

Self-reports are modeled as a function of respondent characteristics X_i and V_i and an error term g_{ri} by the following ordered response equation:

$$Y_{ri}^* = X_i\beta + \varepsilon_{ri};\ \ \varepsilon_{ri} \sim N(0,\sigma_r^2),\ \varepsilon_{ri} \text{ independent of } X_i,\ V_i \tag{4.1}$$

$$Y_{ri} = j \text{ if } \tau_i^{j-1} < Y_{ri}^* \le \tau_i^j,\ \ j = 1,\ldots 5 \tag{4.2}$$

[7] As mentioned earlier, the HRS and PSID have self-report questions on work limiting disabilities on a two-points scale. We will discuss the implications of this below.

The thresholds τ_j^i between the categories are given by

$$\tau_i^0 = -\infty,\ \tau_i^5 = \infty,\ \tau_{\text{i}}^1 = \gamma^1 V_i,\ \tau_i^j = \tau_i^{j-1} + \exp(\gamma^j V_i),\ j = 2,3,4 \qquad (4.3)$$

The fact that different respondents can use different response scales τ_i^j is what we call "differential item functioning" (DIF) (cf. section 4.1).

Using the self-reports on own work disabilities only, the parameters β and γ^1 cannot be separately identified;[8] the reported outcome only depends on these parameters through their difference.

For example, consider country dummies: if two people (with the same characteristics) 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 the self-reports are not enough to identify the work disability difference between the countries.

For example, consider country dummies: if two people (with the same characteristics) 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 the self-reports are not enough to identify the work disability difference between the countries.

Each respondent answered *L=15* vignette questions, five in each of the three domains affect, pain, and heart problems. The evaluations of vignettes *l=1,…,L* are modeled using similar ordered response equations:

$$Y_{li}^* = \theta_l + \theta\,\text{Female}_{li} + \varepsilon_{li} \qquad (4.4)$$

$$Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \le \tau_i^j,\ j = 1,...5 \qquad (4.5)$$

$$\varepsilon_{li} \sim N(0,\sigma^2),\ \text{independent of each other, of } \varepsilon_{ri} \text{ and of } X_i,\ V_i \qquad (4.6)$$

Apart from dummies to indicate the vignettes, the only explanatory variable in (4.4) is a dummy for the gender of the vignette description. The gender dummy is included because preliminary analysis suggested that respondents react differently to vignettes with a female name than with a

[8] The 3 γ^j for *j>1* will still be identified.

male name.[9] The assumption of *"response consistency"* discussed in section 4.1 means that the thresholds τ_i^j are the same for the self-reports and the vignettes.

Given these assumptions, it is clear how the vignette evaluations can be used to 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. The two-step procedure is sketched only to make intuitively clear why the model is identified. In practice, all parameters will be estimated simultaneously by maximum likelihood.[10]

Adjusting for DIF is straightforward in this model once the parameters are estimated. Define a benchmark respondent with characteristics $V_i = V(B)$. (For example, choose one of the countries as the benchmark country.) The DIF adjustment would now involve comparing Y_{ri}^* to the 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 $V(B)$ instead of V_i. Thus a respondent's work ability is computed using the benchmark scale instead of the respondent's own scale. This does not lead to an adjusted score for each individual respondent (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 V_i and X_i. Of course the adjusted distribution will depend upon the chosen benchmark.

4.3. Descriptive Statistics on Vignettes

Based on our estimated models summarized in Table 4, we gave the Dutch and American Internet respondents vignettes in three domains of work disability—pain, affect, and heart disease. The actual vignettes we use in our analysis are presented in Table A2 in the appendix.

Table 6 compares the Dutch evaluations to those in the US. Although the health conditions of the persons described in the vignettes are the same in both countries, there are some substantial differences in the evaluation frequencies. In particular for the two pain and two affect vignettes describing people with relatively mild work limitations, the US respondents

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

[10] This is more efficient than the two-step procedure. Since all error terms are independent, the likelihood contribution is a product of univariate normal probabilities over all vignette evaluations and the self-report, which is relatively easy to compute.

much more often report that these persons have no limitation at all, where the Dutch respondents have a larger tendency to use the intermediate categories “mildly” and “moderately.”

The same tendency towards the extremes in the US and towards the middle for The Netherlands is seen in the fourth vignette, describing a person with relatively serious work limitations (cf. Table A2). The US respondents much more often evaluate this person as severely or extremely limited, where the Dutch still tend to use the answer “moderately.”

Table 6
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	0.2	0.5	0.5
Mildly	63.3	48.9	53.5	46.4	6.2	7.3	7.3	2.6	11.9	8.6
Moderately	10.5	10.9	29.4	21.1	26.6	30.7	31.1	15.4	33.8	38.5
Severely	1.3	0.5	6.27	1.0	50.9	47.1	46.3	58.3	43.9	39.9
Extremely	0.1	1.0	0.30	0.9	16.0	14.7	14.9	23.5	9.9	12.4

Affect vignettes	1		2		3		4		5	
Limited?	NL	US	NL	US	NL	US	NL	US	NL	US
Not at all	32.2	55.1	96.8	97.7	7.4	23.0	12.4	34.2	1.3	8.4
Mildly	54.0	34.1	2.4	0.9	35.3	37.9	43.6	38.4	5.4	11.2
Moderately	11.8	8.7	0.5	0.4	39.7	29.1	31.5	21.3	14.8	20.1
Severely	1.8	1.2	0.3	0.2	16.2	8.7	11.8	5.8	43.3	42.9
Extremely	0.2	0.9	0.1	0.9	1.4	1.2	0.8	0.4	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	9.1	12.9	1.9	3.3	20.5	26.7	7.1	7.5
Mildly	9.8	4.9	49.1	35.7	18.6	15.0	43.3	31.9	36.6	21.2
Moderately	1.0	0.2	28.7	32.7	36.3	32.5	26.2	27.4	31.6	32.4
Severely	0.4	0.0	12.2	16.5	34.3	39.1	9.7	12.4	20.8	30.1
Extremely	0.1	0.9	0.9	2.3	8.9	10.3	0.4	1.7	3.8	8.9

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.

The patterns for the pain and affect vignettes imply that the Dutch seem harder on the vignette persons with a serious limitation and softer on those with a minor limitation. This is the same national tendency for moving away from extremes that we observed in the reports for general health status documented in Table 3. For the two-point scale self-reports on work disability, however, 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 2-point yes/no scale, as people in both categories will be seen as having a work disability by residents of both countries. In contrast, the general reluctance (relative to the Americans) of the Dutch to say that someone is 'not at all' work limited is critical because, as we have seen from the data in Table 1, many of those with mild work limitations are reported to have a work disability on a yes/no scale.

The data in Table 6 suggest that at least in the domains of pain and affect the Dutch would be harder on themselves if they would use the US scales. Using the US scales would thus reduce self-reported work disability prevalence, and would thus also reduce the difference in this prevalence between the two countries. The differences between the countries are much less for vignette evaluations that deal with heart disease—a more 'objective' health condition.

5. 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.[11] 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 limiting disability on a five-point scale (August 2003) and on a two-point scale (October 2003 for 50% of all the observations, December 2003 for the other 50%). Both US surveys have self-reports on 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 in Section 4 with a transformation from the five-point scale to the two-point scale. Table 1 suggested that the cut-off point between "yes" and "no" for the two-point scale is somewhere between the cut-off points between "no" and "mildly" and "mildly" and "moderately" for the five-point scale. In line with this, 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:

$$\tau_i(2) = \lambda\tau_i^1 + (1-\lambda)\tau_i^2 \qquad (5.1)$$

[11] 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.

We assume that the weight λ does not vary with individual characteristics and is the same in the US and The Netherlands. 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. If the Dutch have lower thresholds on the five-point scale, they also have a lower threshold on the two-point scale, etc. This assumption is needed since there are no five-point scale self-reports for the US. The parameter λ is identified from the Dutch self-reports on both scales (and then also applied to the US respondents). All parameters are estimated simultaneously by Maximum Likelihood, taking into account that for the US respondents, the five-point scale self-report is not observed.

Table 1 suggests that there is some random error in the two-point and/or five-point scale evaluations that is not transferred to the other scale. To account for this, we adjust the equation for the respondent's own work limiting disability by partitioning the error term in a genuine unobserved component of work disability affecting both the two-point and the five-point scale reports, and an idiosyncratic error term affecting only one report and independent of everything else. To be precise, the two-point scale and five-point scale self-reports are modeled as follows:

$$Y_{ri}^* = X_i\beta + \varepsilon_{ri};\ \ \varepsilon_{ri} \sim N(0,\sigma_r^2),\ \varepsilon_{ri} \text{ independent of } X_i,\ V_i \qquad (5.2)$$

Five-point scale:

$$Y_{ri}^5 = j \text{ if } \tau_i^{j-1} < Y_{ri}^* + u_i^5 \le \tau_i^j,\ \ j = 1,\ldots 5 \qquad (5.3)$$

Two-point scale:

$$Y_{ri}^2 = 0 \text{ if } Y_{ri}^* + u_i^2 \le \tau_i\ (2);\ \ Y_{ri}^2 = 1 \text{ if } Y_{ri}^* + u_i^2 > \tau_i\ (2) \qquad (5.4)$$

$$\begin{aligned} &u_i^2, \sim N(0,\sigma_{u^2}^2); u_i^5 \sim N(0,\sigma_{u^5}^5); \\ &u_i^2, u_i^5 \text{ independent of each other and of other error terms} \end{aligned} \qquad (5.5)$$

The equations for work disability and for the thresholds all include a complete set of interactions with the country dummy for The Netherlands. Vignette evaluation equations and the auxiliary parameters introduced above concerning the transformation from the two-point to the five-point scale do not include such interactions.

5.2 Within Country Implications

Within this basic structure we consider a variety of models in order to test the sensitivity of our main results to different assumptions. The first model, which we term the "benchmark model" uses all 15 vignettes covering the three domains (affect, pain and cardiovascular disease), assuming a common response scale across these domains.

Table 7 presents the results for the work disability equation in this model, comparing it with a model that does not allow for any threshold variation across respondents. The latter model (first two columns) essentially reproduces the probits presented in Section 3, the difference being that for estimation, all age groups are now combined. The remaining columns illustrate the effects of allowing for different response scales. The middle two columns represent the estimated effect of respondent characteristics on the first response threshold γ^1, which is the critical threshold for determining whether someone claims to be work disabled on a two point scale[12]. The final two columns are the coefficients in the work disability equation after correcting for differential response scales. The model that does not allow for response scale variation is strongly rejected against the more general model that does allow for DIF. The same result is found for each country separately, in line with the many significant parameters in the first threshold in Table 7.

The biggest adjustment relates to the principal focus of this paper—the different response scales used by the Dutch and the Americans. This corresponds to the large negative coefficient on the dummy for the Netherlands in the first threshold, implying that the Dutch use lower thresholds. We explore the across country difference in more detail below, but first focus on the impact of accounting for differential response scales on conclusions on within country variation in work disability.

In the model without the DIF correction in the US, work disability 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 the education effects are steeper in the US than in the Netherlands. The vignette corrected results for the US imply an even larger fall in reported work disability across education groups, since significantly higher thresholds for work disability are used by those in the lowest

[12] For reasons of space we do not report the estimates for γ_2, γ_3, and γ_4.

education category compared to the higher educated groupsn (γ^1). 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.

Table 7
Estimation Results Benchmark Model—Work Disability

	Model without DIF		Complete Model			
	Work disability		First Threshold Parameter		Work disability	
	β	s.e.	γ^1	s.e.	β	s.e.
Constant	-40.66	8.07*	0.00	0.00	-41.19	8.20*
Ed med	-3.19	0.30*	-0.76	0.20*	-3.88	0.34*
Ed high	-5.32	0.42*	-0.80	0.25*	-5.96	0.47*
Age/100	78.61	25.76*	-5.83	9.39	72.64	27.51*
(Age/100)^2	-46.71	20.47*	-1.95	8.14	-48.45	22.06*
Female	-0.28	0.26	1.23	0.17*	0.82	0.30*
Hypertension	2.04	0.27*	0.39	0.18*	2.36	0.31*
Diabetes	4.03	0.36*	-1.21	0.31*	2.73	0.45*
Cancer	2.63	0.41*	0.03	0.28	2.64	0.47*
Disease of lung	5.95	0.44*	0.98	0.32*	6.78	0.52*
Heart problem	5.64	0.35*	0.26	0.39	5.86	0.47*
Emotion	6.18	0.39*	-2.23	0.29*	4.05	0.47*
Pain	10.62	0.39*	-0.55	0.17*	10.25	0.44*
Interactions with dummy NL						
Constant	14.19	8.90	-7.89	2.67*	6.94	9.37
Ed med	2.81	0.86*	0.57	0.21*	3.35	0.89*
Ed high	2.46	0.94*	1.01	0.26*	3.34	0.97*
Age/100	-26.97	29.63	6.48	9.47	-21.31	31.28
(Age/100)^2	5.17	24.78	0.51	8.23	6.56	26.22
Female	1.28	0.73	-1.45	0.18*	0.07	0.76
Hypertension	-1.29	0.87	-0.39	0.20	-1.47	0.90
Diabetes	1.52	1.60	-0.12	0.36	2.01	1.67
Cancer	-0.28	1.49	-0.02	0.35	-0.12	1.53
Disease of lung	0.92	1.33	-1.03	0.37*	0.12	1.40
Heart problem	2.88	1.25*	0.12	0.41	2.93	1.33*
Emotion	1.88	0.99	1.10	0.30*	3.07	1.07*
Pain	4.83	0.84*	1.05	0.20*	5.63	0.89*

Normalization: $\sigma_r^2 = 10$; * : significant at two-sided 5% level.

In the US, we estimate that women use higher thresholds than men. The results correcting for this show that women actually have a significantly larger probability to be work disabled, given health conditions and demographics. In the Netherlands, there is no evidence that male and female respondents use different thresholds. After correcting for DIF, the effect of gender on work disability is almost the same in both countries.

In addition to the use of different thresholds by gender, we also find evidence that the threshold used was different if a female name was used in the vignette discription instead of a male name (the parameter θ in equation (4.4)). We find that, 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.[13]

Pain, emotional problems, and heart problems are more important causes for work disability in the Netherlands than in the US. Correcting for DIF increases the difference between the effects of these variables in the two countries, particularly for emotional problems. In both countries, we find that respondents with emotional problems tend to use lower thresholds than respondents without emotional problems, but the difference is larger in the US than in the Netherlands.

As a consequence, the effect of emotional problems on work disability is overestimated in the estimates not correcting for DIF, particularly in the US. Correcting for this increases the difference between the effects in the Netherlands and the US, which was already positive (but insignificant) in the model without DIF. For pain, the result is slightly different. In the US, people with pain use lower thresholds than people without pain, but in the Netherlands, we find the opposite, implying that the effect of pain on work disability in the Netherlands is underestimated in the model without DIF.

5.3 Implications for Across Country Comparisons: Benchmark Model and Variants

In this section and the next, we present simulations based on our models to address 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 the CentERpanel to make the samples population representative of the 51-64 age groups in the two

[13] Interactions of respondent gender and gender of the vignette person were insignificant.

countries. Our simulations take the explanatory variables in the two samples as given and simulate values of work disability using US thresholds for both the US and the Dutch sample.

Our simulations start with the model without DIF and the benchmark model, the parameters of which were discussed in the previous section. We will also consider several alternative models, summarizing the five-point answers in three categories, not including health conditions as regressors, or relaxing the assumption of common response scales across domains.

Table 8 compares predictions of work disability on the two-point US response scale of the various models. The first line refers to the model with the same response scales for all respondents in the US and The Netherlands, cf. the first two columns in Table 7. This line approximately reproduces the difference in work disability between the two countries for the age group 51-64, with work disability in the Netherlands about 57% larger than in the US. The second line presents simulations based on the benchmark model. If we give the Dutch respondents the response scales of their US counterparts with the same age, education level, gender and health conditions, the predicted disability rate in The Netherlands among 51-64 year olds drops from 35.8% to 28.0%. Equivalently, the percent difference in work disability between The Netherlands and the US drops from 57.5% to 23.4%. This implies that more than half of the cross-country difference in the work disability rate in the two countries is explained by response scale differences.

Since the threshold on the two-point scale is a weighted average of the first two thresholds on the five-point scale (with estimated weights of 0.79 for the first threshold and 0.21 for the second threshold according to the benchmark model), the third and fourth threshold of the five-point scale seem unimportant for predicting work disability at the two-point scale. Therefore, combining the three categories moderate, severe and extreme work disability should not have a large effect on the results. To check this intuition, we reestimated the model after combining these three categories—without the parameter vectors γ^3 and γ^4 that determine the third and fourth threshold. The third line of Table 8 presents results for the benchmark model in which the three categories moderate, severe and extreme work disability are combined for both the five-point scale self-reports and the vignette evaluations. The estimated effect of the response scale differences between the two countries is somewhat larger still than in the benchmark model, and the estimated difference in work disability according to US response scales goes down to 15.3%.

Table 8
Predicted Work Disability Age Group 51-64 using US Response Scales—Several Models

	Percentage Work Disabled		
	NL	US	% Difference NL-US
Model without DIF	35.76	22.71	57.5%
Benchmark model using all vignettes	28.02	22.71	23.4%
Model combining moderate, severe, extreme	26.24	22.76	15.3%
Model not using health conditions	27.49	23.04	19.3%
Model using affect vignettes only	22.39	22.76	-1.6%
Model using pain vignettes only	28.30	22.77	24.3%
Model using cvd vignettes only	31.95	22.77	40.3%

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

The second issue for our sensitivity analysis is the use of health conditions as regressors. Until now, we have ignored potential measurement problems with these reported health conditions. Baker, Stabile and Deri (2004) show that this assumption may be problematic. Particularly if there are systematic differences in reporting health conditions across countries, this might bias our results. We have therefore also reestimated the benchmark model without using any information on health conditions, excluding the health conditions from the equations for work disability and from the equations for the thresholds. The results in the fourth row of Table 8 shows 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 19.3%, not that far from the 23.4% in the benchmark model.

Until now, we have assumed 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 reestimated the benchmark model using the vignettes in only one of the three domains.

The resulting predictions are presented in the final three rows of Table 8. 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: if everyone would use the US response scales, the work disability rate in the US would be slightly larger than in the Netherlands (0.4%-points—or 1.6%). But if we would only use the vignettes on heart problems, the opposite conclusion is obtained: respondents in the US and The Netherlands use similar response scales, and only a small part of the difference in self-reported work disability rates is explained by response scale differences. When using the US response scales in both countries, the cvd vignettes based estimates still give a 9.2%-points difference between The Netherlands and the US. For the pain vignettes, the results are in between these two extremes and similar to those for the benchmark model. These results cast doubt on the assumption of common response scales across health domains, which motivates an alternative model that is more general in the sense that it accounts for different response scale differences for the various domains.

5.4 Implications for Across Country Differences: A General Model

In this model it is assumed that true work limitations are the maximum of work limitations in different health domains. The domains are Affect, Pain, CVD, and "Other". For the former three domains we have vignettes, which can be used to correct scale differences across the two countries, but not for "Other." For Affect, Pain and CVD, we assume that only respondents who have the corresponding health problem can report work disability due to a problem in that domain. For example, the equation for work disability in the affect domain only applies to respondents reporting that the doctor has ever told them that they have an emotional health problem. The details of the model are outlined in Appendix 1.

Here, we only discuss some simulation results from this model, presented in Table 9. The format of this table is similar to that used in Table 4. The first panel gives the predictions for the age group 51-64 if everyone in each country uses their own response scales. For example, in The Netherlands, about 47% of those with an emotional condition would classify themselves as work disabled, versus only 27% in the US.[14] The second panel shows that this difference is almost completely due to response scale differences: if the Dutch respondents would use the (higher) US response scales, then 27% of the Dutch with an emotional health condition would report themselves as disabled, very similar to the predicted rate in the US.

[14] The observed (not domain specific) work disability rates in the age group 51-64 (weighted with sample weights) among those with an emotional health condition are 51.7% in the US and 76.2% in the Netherlands. These numbers are larger than the simulated rates in the text since the latter refer to (affect) domain specific work disability.

Table 9

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

Panel 1- Predictions using own response scales:

Domain	Work disability in group with health condition		Prevalence of health condition		Work disability in population	
	NL	US	NL	US	NL	US
Affect	47.2	27.1	10.5	14.3	5.0	3.9
Pain	65.6	36.1	33.7	27.6	22.1	10.0
CVD	41.7	26.7	12.2	15.6	5.1	4.2
a,p,c	62.9	37.7	46.1	42.0	29.0	15.8
Other					12.4	9.4
Total					37.2	23.1

Panel 2-Predictions using US response scales

Domain	Work disability in group with health condition		Prevalence of health condition		Work disability in population	
	NL	US	NL	US	NL	US
Affect	27.3	27.1	10.5	14.3	2.9	3.9
Pain	51.5	36.1	33.7	27.6	17.4	10.0
CVD	37.1	26.7	12.2	15.6	4.5	4.2
a,p,c	49.2	37.7	46.1	42.0	22.7	15.8
Other scale = affect scale:						
Other					7.7	9.4
Total					27.9	23.1
Other scale = pain scale:						
Other					9.9	9.4
Total					29.8	23.1
Other scale = cvd scale:						
Other					12.5	9.4
Total					31.4	23.1

Notes: CentER Savings Survey 2003 for The Netherlands and HRS 1998 for the US, weighted with sampling weights.

Multiplying these numbers by the prevalence rates of emotional health problems (in the middle panel) gives work disability in the emotional health domain as a percentage of the complete age group. Once response scale differences are adjusted for, this is very similar in the two countries.

The results for work disability in the domains of pain and heart problems are quite different. The prevalence rate for pain is smaller in the US than in The Netherlands, as we saw before. In The Netherlands, the probability that people who often have pain would report a pain related work disability is almost twice as large as in the US.[15] While the difference would be a lot smaller if the Dutch would use the US response scales, it would not disappear. Even then, work disability in the pain domain would explain a more than 17% work disability rate in The Netherlands compared to 10% in the US.

For heart problems, the response scales in the two countries are rather similar, so that there is only a small adjustment if response scale differences are controlled for. US respondents more often report that the doctor has told them that they have a heart problem than Dutch respondents, but Dutch respondents with heart problems have a substantially larger probability to be work disabled.[16] Since the latter difference is larger than the former, the rate of heart problems related work disability is somewhat larger in The Netherlands than in the US.

Comparing the three domains, we find that there is more pain related work disability than affect or cvd related work disability in both countries. For the US, this is at least qualitatively in line with the HRS data on the most important source of work disability—the most common reported sources are back, neck and spine problems. In the US, work disability due to heart problems and emotional problems are about equally likely. In The Netherlands, work disability due to emotional problems would be more common than work disability due to heart problems if the Dutch response scales are used, but this reverses using the US response scales.The fourth row in each panel (labeled a,p,c) shows how many respondents suffer from work disability in at least one of the three domains. In the US, these three domains already give a work disability rate of 15.8%, 69% of the total work disability rate in this age group. In The Netherlands, and using Dutch response scales, the three domains explain even almost 78% of total work disability. Combining the three domains, the difference in work disability in either of these three domains between The Netherlands and the US reduces from 13.2%-points to 6.9%-points if response

[15] Observed (not domain specific) work disability rates among those with pain are 51.0% in the US and 74.5% in the Netherlands.

[16] This difference corresponds to the observed (not domain specific) work disability rates among respondents who report a heart condition: 49.9% in the US and 65.4% in the Netherlands.

scale differences are adjusted for. Thus about half of the gap is due to response scale differences, a conclusion similar to that based upon the benchmark model.

Work disability rates due to other health problems than heart problems, emotional problems, or pain, are 12.4% in The Netherlands with Dutch response scales, and 9.4% in the US with US response scales, and part of the difference could be due to different response scales. Combining *Other* with the three domains affect, pain, and cvd then gives total work disability rates using country specific response scales of 37.2% and 23.1%, close to the work disability rates in the raw data. For this domain, we cannot correct for response scale differences in the same way as for the other three domains, since no vignettes on *Other* are available.

In the second panel of Table 9, we have assumed that DIF for *Other* is equal to DIF for either affect, pain, or cvd. Accordingly, larger or smaller corrections for work disability are obtained. If we assume that the response scale difference for *Other* is the same as that for affect, then work disability in the *Other* domain in The Netherlands on US response scales would fall to 7.7%, lower than the 9.4% in the US. Estimated total work disability in The Netherlands on US response scales would then be 27.9%. Adjusting for response scale differences would thus reduce the difference in overall work disability between The Netherlands and the US from 14.2%-points to 4.8%-points. If the response scale difference for *Other* equals that for pain or cvd, then work disability in the other domain in the Netherlands on the US scale is larger – 9.9% or 12.5% - and total work disability would be 29.8% or 31.4%, reducing the gap with the US to 6.7%-points (using the pain scales difference for other) or 8.3%-points (using the cvd scales difference for other). Depending on which correction is used for *Other*, we can therefore conclude from this general model that response scale differences explain between 42% and 66% of the gap in self-reported work disability rates in the two countries for the 51-64 age group.

6. Conclusions

Workers in different western countries report very different rates of work disability**.** This diversity in reported work disability stands in sharp contrast to the believed similarity in their health outcomes. This contradiction continues to be seen as a major unresolved puzzle.

In this paper, using new data from two of these countries- the US and The Netherlands- we offer a partial resolution of the puzzle. Our resolution claims that a significant part of the observed difference in reported work disability between the countries lies in the fact that residents of the two countries use different response scales in answering the 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.

We were able to reach this conclusion by implementing a vignette methodology into Internet surveys, which we conducted in both countries. Our vignettes gave respondents in both countries 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 important and more subjective health domains of pain and emotion, the evidence is quite 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, based on this research there seems little question that they exist. While one may quarrel with the specific assumptions in each of our modelling approaches outlined 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 can be a useful tool in helping us understand within country differences in reporting as well. For example, 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.

Vignettes represent a potentially important new methodological tool that may aid in the analyses of other applications besides health and disability. Any time treshold scales are used to categorize individual responses, the question will arise on whether people really differ in their response 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 in certain domains, consumer satisfaction, measurement of risk, and perception of poverty.

The application of a new technique like the use of vignettes poses new methodological and empirical questions. Internet surveys appear to be a quite powerful tool to address such questions.

References

Aarts, L., R. Burkhauser and P. De Jong. 1996. *Curing the Dutch disease: an international perspective on disability policy reform,* Aldershot, Avebury.

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

Baker, M., M. Stabile and C. Deri. 2004. What do self-reported, objective measures of health measure?, *Journal of Human Resources*, forthcoming.

Bound, J. and R. Burkhauser. 1999. Economic Analysis of Transfer Programs Targeted on People with Disabilities, *Handbook of Labor Economics, Vol. 3C*, O. Ashenfelter and D. Card (eds.), 3417-3528.

Bound, J. 1991. Self-reported versus objective measures of health in retirement models, *Journal of Human Resources,* 26(1), 106-138.

Burkhauser, R. and M. Daly. 2002. Policy Watch: U.S. disability policy in a changing environment, *Journal of Economic Perspectives,* 16(1), 213-224.

Burkhauser, R., M. Daly, A. Houtenville and N. Nargis. 2002. Self-Reported Work Limitation Data—What They Can and Cannot Tell Us, *Demography*, 39(3), 541-555.

Currie, J. and B. Madrian. 1999. Health, health insurance and the labor market, *Handbook of Labor Economics, Vol. 3C*, O. Ashenfelter and D. Card (eds.), 3309-3416.

DeLeire, T. 2000. The wage and employment effects of the American with Disabilities Act, *Journal of Human Resources,* 35(4), 693-715.

Eurostat. 2001. *Disability and Social Participation in Europe*, Luxembourg: Office for Official Publications of the European Communities.

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

Kerkhofs, M. and M. Lindeboom. 1995. Subjective health measures and state department reporting errors, *Health Economics*, 4, 221-235.

King, G., C. Murray, J. Salomon, and A. Tandon. 2004. Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research, *American Political Science Review* 98(1), 567-583.

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

Stapleton, D. C. and R. V. Burkhauser (eds.). 2003. *The Decline in Employment of People with Disabilities: A Policy Puzzle*. Kalamazoo, MI: W.E. UpJohn Institute for Employment Research.

Appendix: General Model with Multiple Domains of Work-Releated Health

Let the health domains determining work related health be given by $d=1,...,D$. In the empirical work, we will use $D=4$, with domains *affect, pain, heart problems*, and *other*. For the first three *(d=1,2,3),* we assume that only those who report a health condition in that domain can suffer from a work disability in that domain. Such an assumption is not made for the domain *other (d=4),* since we do not want to assume a priori that the health conditions observed in the survey are a complete description of all health conditions that could possibly lead to a work related health problem. (And indeed, the raw data contain people who report a work related disability while they do not report to have any of the observed health conditions.)

Respondent work limitations due to problems in dimension *d* are given by:

$$Y_{ri}^{*}(d) = \beta(d)' X_i + \varepsilon_{ri}(d) \qquad (1.1)$$

For $d=4$, this equation applies to all respondents; for $d=1,2,3$, it only applies to those who report the corresponding health condition; for the others, $Y_{ri}*(d)$ will be minus infinity. Response scales can vary across domains. The response scale in dimension *d* will be given by

$$\tau_i^{\,0}(d) = -\infty,\ \tau_i^{\,3}(d) = \infty,\ \tau_i^{1}(d)= \gamma^{\,1}(d)Vi;\ \tau_i^{\,2}(d) = \tau_i^{\,1}(d) + exp(\gamma^{\,2}(d)X_i).$$

Here we have merged the categories *moderately limited*, *severely limited* and *extremely limited/cannot work* into one category to reduce the total number of parameters to be estimated, reducing the five-point scale to a three-point scale (cf. the third model in Table 8).

If work limitations due to health problems in domain *d* were asked, they would be reported as

$$Y_{ri}\ (d) = j \text{ if } \tau_i^{j-1}(d) < Y_{rd}^{*} \leq \tau_i^{j}(d), \quad j=1,...,3,\ d=1,...,D\ (=4)$$

In the available data, however, work limitations in specific domains are not reported.[17] The only question is whether there is any health problem that leads to work limitations. It seems reasonable to interpret the answer to this as the maximum of the work limitations in all domains:

$$Y_{ri} = \max\{Y_{ri}(1),...,Y_{ri}(D)\}$$

[17] The HRS asks people reporting some work disability, which domain(s) cause(s) the work disability. We do not use this information in the formal models, since the categories do not match our domains and since no such information is available for the Dutch data.

To identify the model even in the standard case of no variation in response scales, some assumptions are needed to distinguish the four determinants of reported overall work disability. The three domains *affect, pain* and *heart problems*, clearly relate to reported health conditions, emotional problems (has the doctor ever told you that you have emotional, nervous or psychiatric problems?), pain (do you often have pain?) and heart problems (has the doctor ever told you that you had a heart attack, coronary heart disease, angina, congestive heart failure, or other heart problems? Or has the doctor ever told you that you had a stroke or transient ischemic attack?). As already explained above, we will assume that respondents can only report a work disability[18] in one of these three domains if they suffer from that type of health condition. The health condition dummies are included as exogenous variables in our model (as before). Thus this implies that for someone who reports none of these three health conditions, only the domain *other* can lead to work related disability. On the other hand, respondents who have an emotional problem but no heart condition and who do not suffer from pain can be work disabled in either the *affect* domain or in the *other* domain (or both).

Moreover, we will assume that work related health in the three domains *affect, pain* and *heart problems* is not affected by other health conditions. Thus having diabetes, cancer, a lung disease, arthritis, or hypertension can only lead to work disability through the *other* domain. This implies zero restrictions on $\beta(1)$, $\beta(2)$ *and* $\beta(3)$—all coefficients corresponding to the health conditions are set to zero (since the equations only apply to those with the given health condition, this also applies to the corresponding health condition itself—the intercept and the coefficient on that health condition capture the same thing).

No restrictions are imposed on $\gamma(1)$-$\gamma(3)$ (except the normalization on the constant term: as in the benchmark model, this is set equal to zero for each domain)*;* these are identified by the vignettes in these three domains. Since here are no vignettes on the domain *other,* $\gamma^1(4)$ is not identified. We consider alternative assumptions on this parameter vector in the simulations (cf. Table 9).

Without vignettes, these assumptions are sufficient to identify $\beta(d)-\gamma^1(d)$, *d=1,…,D,* but not the parameters of interest $\beta(d)$, *d=1,…,D*. Vignettes can be used to identify the parameters for the domains for which vignettes are available, in our case *d=1,2,3*.

[18] That is, report a 'yes' on the two-point scale or report a mild limitation or worse on the five point scale.

The vignette descriptions explicitly refer to problems in one domain, stating that the vignette-persons have no other health problems. Thus for the vignettes in domain *d*, it is reasonable to assume that work limitations in dimension *d* are larger than work limitations in other dimensions and completely determine the answer to the vignette work limitations question. This gives the following model for observed vignette evaluations, $Y_l(d), l=1,\ldots,L$ (*L* vignette descriptions for each dimension ; $L=5$ in our case), $d=1,\ldots D^v$ ($1,\ldots,D^v$ are the dimensions for which vignette descriptions are available; 3 in our case).

$$Y_{li}^*(d) = \theta_l(d)+\psi_l(d)\ Female_{li} + \varepsilon_{li}\ (d); \qquad Y_{li}\ (d) = j \text{ if } \tau_i^{j-1}(d) < Y_{li}^*(d) \leq \tau_i^j(d), \quad j=1,\ldots,K$$

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

The vignette reports identify $\gamma^1(d)$ except for the constant terms $(d=1,2,3)$ and $\theta_l(d)$, $l=1,\ldots,5;\ d=1,2,3,$ up to a constant term for each domain and $\gamma^j(d)$ for $j>1$. The self-reports then identify $\beta(d)$. This is the same "correction" that was carried out in the one-dimensional case.

To estimate the model, an assumption needs to be made on the joint distribution of the errors. We assume joint normality and independence of each other and of the (thus exogenous) variables X_i.

The assumptions on the relation between the two-point scale and the five-point scale remain the same as before. We also assume that this relation is the same for all domains.

Finally, we list the parameters in each equation of the multi-domain model :

- Respondent work disability in domains $d=1,2,3$: equation includes intercept, 5 demographics, and 6 interactions with the NL country dummy. The variance of the error term is normalized at 100 (fixing the scale). This gives 36 parameters. See the results in Table A3.
- Response scales (2 thresholds) in domains $d=1,2,3$: demographics plus health conditions other than the one corresponding to this particular disability, with all interactions with dummy NL. To normalize location: no intercept in threshold 1. This gives 3*(23+24)=141 parameters.
- Respondent work disability in domain 4 *(other)*: intercept, 5 demographics, 4 health conditions, interactions with dummy NL. Error term has variance 100. This gives 20 parameters. See the results in Table A3.
- Respondent work disability threshold 1 *other*: not identified.

- Respondent work disability threshold 2 *other*: only identified for NL (since there are no 5-point scale answers in the US on this domain, neither self-reports nor vignettes); 10 parameters.
- Vignette dummies, coefficients on gender of the vignette persons, standard deviations of vignettes. 3*(5+1+1)=21 parameters.
- Three auxiliary parameters transforming the five-point scale into the two-point scale (two standard deviations of idiosyncratic noise (independent across the two scales) and one for the weight of threshold 1; all assumed the same across domains and countries).

In total: 36+147+20+10+21+3=237 parameters.

Table A1: Probits for Work Disability, Age Bracket 51-64

	Netherlands		U.S.	
	Coefficient	DF/dX	Coefficient	DF/dX
High Blood Pressure	0.042	0.015	0.154	0.043
	(0.26)	(0.26)	(4.60)**	(4.60)**
Diabetes	0.107	0.038	0.388	0.120
	(0.35)	(0.35)	(8.31)**	(8.31)**
Cancer	0.069	0.025	0.241	0.072
	(0.20)	(0.20)	(4.07)**	(4.07)**
Lung Disease	1.358	0.500	0.440	0.140
	(3.38)**	(3.38)**	(7.47)**	(7.47)**
Heart Problems	0.715	0.274	0.504	0.159
	(2.63)**	(2.63)**	(11.35)**	(11.35)**
Stroke	1.817	0.602	0.850	0.298
	(2.79)**	(2.79)**	(9.98)**	(9.98)**
Arthritis	0.612	0.232	0.353	0.099
	(2.64)**	(2.64)**	(10.30)**	(10.30)**
Emotional Problems	1.321	0.491	0.608	0.196
	(5.47)**	(5.47)**	(14.02)**	(14.02)**
Pain	1.600	0.571	0.917	0.290
	(9.80)**	(9.80)**	(26.26)**	(26.26)**
Female	0.100	0.035	-0.072	-0.020
	(0.67)	(0.67)	(2.15)*	(2.15)*
Ed_med	0.054	0.019	-0.327	-0.092
	(0.31)	(0.31)	(8.95)**	(8.95)**
Ed_hig	-0.066	-0.023	-0.567	-0.134
	(0.38)	(0.38)	(10.76)**	(10.76)**
Constant	-1.396		-1.309	
	(8.32)**		(30.31)**	
Observations	503		9393	
Observed p	0.34		0.25	
Log Likelihood	-202.76		-3814.78	

Robust z statistics in parentheses
* significant at 5%; ** significant at 1%

Table A2: Vignette Descriptions Used in Both the U.S. and The Netherlands
(All vignettes are presented with either a female or a male name, which are randomized across respondents. Here we only show one of the two names per vignette)

Vignettes for Affect

1. [Henriette] generally enjoys her work. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day-to-day activities on the job.

2. [Jim] enjoys work very much. He feels that he is doing a very good job and is optimistic about the future.

3. [Tamara] has mood swings on the job. When she gets depressed, everything she does at work is an effort for her and she no longer enjoys her usual activities at work. These mood swings are not predictable and occur two or three times during a month.

4. [Eva] feels worried all the time. She gets depressed once a week at work for a couple of days in a row, thinking about what could go wrong and that her boss will disapprove of her condition. But she is able to come out of this mood if she concentrates on something else.

5. [Roberta] feels depressed most of the time. She weeps frequently at work and feels hopeless about the future. She feels that she has become a burden to her co-workers and that she would be better dead.

Vignettes for Pain

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. [Catherine] suffers from back pain that causes stiffness in her back especially at work but is relieved with low doses of medication. She does not have any pains other than this generalized discomfort.

3. [Yvonne] has almost constant pain in her back and this sometimes prevents her from doing her work.

4. [Jim] has back pain that makes changes in body position while he is working very uncomfortable. He is unable to stand or sit for more than half an hour. Medicines decrease the pain a little, but it is there all the time and interferes with his ability to carry out even day to day tasks at work.

5. [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.

Vignettes for CVD

1. [Trish] is very active and fit. She takes aerobic classes 3 times a week. Her job is not physically demanding, but sometimes a little stressful.

2. [Norbert] has had heart problems in the past and he has been told to watch his cholesterol level. Sometimes if he feels stressed at work he feels pain in his chest and occasionally in his arms.

3. [Paul]'s family has a history of heart problems. His father died of a heart attack when Paul was still very young. The doctors have told Paul that he is at severe risk of having a serious heart attack himself and that he should avoid strenuous physical activity or stress. His work is sedentary, but he frequently has to meet strict deadlines, which adds considerable pressure to his job. He sometimes feels severe pain in chest and arms, and suffers from dizziness, fainting, sweating, nausea or shortness of breath

4. [Tom] has been diagnosed with high blood pressure. His blood pressure goes up quickly if he feels under stress. Tom does not exercise much and is overweight. His job is not physically demanding, but sometimes it can be hectic. He does not get along with his boss very well.

5. [Dan] has undergone triple bypass heart surgery. He is a heavy smoker and still experiences severe chest pain sometimes. His job does not involve heavy physical demands, but sometimes at work he experiences dizzy spells and chest pain.

Table A3: Work Disability Respondent in Several Domains

	Affect		Pain		Heart Problems		Other	
	par.	s.e.	par.	s.e.	par.	s.e	par.	s.e.
Constant	-65.175	43.622#	-67.078	21.086*	-30.179	36.401	-20.111	11.685+
Ed_med	-4.698	1.318*	-3.764	0.901*	-3.480	1.228*	-3.180	0.680*
Ed_high	-8.676	1.824*	-7.018	1.279*	-6.055	1.702*	-4.694	0.843*
Age /10	22.461	14.216#	21.451	7.122*	18.839	12.050#	0.030	3.815
(Age/10)^2	-1.917	1.149+	-1.671	0.578*	-1.673	0.958+	0.239	0.311
Woman	-1.151	1.272	1.898	0.783*	-0.313	1.131	0.463	0.592
High blood							1.254	0.610*
Diabetes							10.312	2.482*
Cancer							3.722	0.917*
Lung							9.757	1.058*
Interactions with dummy NL								
Constant	37.268	47.600	42.116	23.538+	44.872	51.983	3.670	12.361
Ed_med	-3.268	3.275	7.011	1.931*	-1.917	4.206	2.022	1.270#
Ed_high	-3.181	3.802	7.070	2.358*	-1.126	4.896	2.144	1.331#
Age /10	-13.776	16.097	-11.289	7.894#	-16.109	16.543	0.224	4.144
(Age/10)^2	1.236	1.377	0.757	0.664	1.507	1.317	-0.152	0.349
Woman	2.741	3.037	-2.604	1.674#	6.660	4.162#	0.962	1.049
High blood							0.983	1.251
Diabetes							0.146	2.489
Cancer							1.733	2.157
Lung							1.236	2.211

Notes: Normalization: $\sigma_r^2 = 10$.

*Significant at (two-sided) 5% level; + significant at 10% level; # significant at 20% level.